0751

10-12-83

Verona Plant Fish and Sediment Plan

Submitted by
Syntex Agribusiness, Inc.
October 12, 1983

As Revised
January 31, 1984
and
March 9, 1984

8403.02
8351G/465

FISH AND SEDIMENT PLAN

I. INTRODUCTION

Syntex Agribusiness, Inc. ("Syntex") has entered into a Consent Agreement and Order with the United States Environmental Protection Agency ("EPA"). This Consent Agreement and Order provides, among other things, that Syntex shall develop and submit to EPA for approval a plan ("Fish and Sediment Plan") for the sampling and analysis of 2,3,7,8-tetrachlorodibenzo-p-dioxin ("dioxin" or "TCDD") in fish, and TCDD in sediment in the Spring River at selected locations downstream from Syntex' Verona, Missouri facility (the "Facility"). The Fish and Sediment Plan initially provides for the sampling and analysis of Spring River fish and sediment for a five (5) year period extending up to twelve (12) miles downstream from the Facility. Such period and/or distance may be extended or shortened by mutual agreement or based on the results obtained. The Fish and Sediment Plan includes a discussion of the sampling locations, analytical procedures, statistical methodology and a schedule of implementation.

II. SAMPLES

A. Fish. Samples of fish will initially be obtained annually for a period of five years unless shortened by mutual

agreement of Syntex and EPA based on the results obtained. The samples will be obtained by the Missouri Department of Conservation ("MDC") within the period August 1 to August 31 and will consist of twenty bottom feeding fish taken from each of the locations described below. The size and weight of the fish samples will be as consistent as possible from year to year. The weight and length of each fish will be recorded by MDC in the sampling log ("Sampling Log"), using the format which may be found at Attachment D. Approximate sampling locations are designated by an "o" on the Spring River map at Attachment A. Additional information concerning these locations may be found at Attachment H. It is intended that the samples be taken at the following sampling locations (or as near thereto as access to the river permits):

(1) 0.3 miles downstream from the Facility ("Location 1");

(2) 3.0 miles downstream from the Facility ("Location 2");

(3) 6.0 miles downstream from the Facility ("Location 3");

(4) 9.0 miles downstream at road H near Hoberg ("Location 4");

(5) 12.0 miles downstream near intersection with road V ("Location 5").

Sampling locations, in any event, shall be the same from year to year.

MDC will be responsible for obtaining fish by electroshock and for classifying and labeling the fish samples at each location. EPA will provide Syntex, along with the samples, a brief description of the methods used to gather and store them and such methods shall be consistent, in material respects, from year to year. Additionally, sufficient fish shall be gathered by MDC such that, to the extent practicable, comparable species of fish may be analyzed from each location from year to year. MDC will rank the twenty fish from each of the locations according to size, and will sequentially divide them into two groups for each location ("Groups") so as to provide comparable size representation. All fish samples will then be promptly filleted without skin by MDC (at its Fish and Wildlife Center on 1110 College Avenue, Columbia, Maryland 65201), using the "standard fillet procedure", a copy of which may be found at Attachment E. The fillets from each of the two Groups for each location will then be weighed, homogenized and analyzed as two separate samples. The remaining fish parts from one of the two Groups for each location will be weighed, homogenized and analyzed as one sample. That result will then be used with the result from the corresponding fillet analysis for the Group to compute, by weighted average, a "whole fish

body" value and will then be frozen and stored by MDC prior to their shipment by EPA to an appropriate analytical laboratory designated by Syntex.

An additional sample consisting of four to ten bottom feeders will be taken for whole fish analysis at Location 1 to allow further interpretation of existing whole-body residue data from the Spring River. The data from these whole fish samples will not constitute part of the statistical analysis described in Section V.

The completed Sampling Log will be supplied by MDC to EPA for transmittal to Syntex. Upon receiving notification from EPA that the samples are ready, Syntex will confirm the identity of the designated laboratory and EPA will be responsible for delivery of the samples to such laboratory for analysis. It is currently anticipated that these samples will be analyzed at the University of Nebraska laboratory under the supervision of Dr. Michael Gross, at Syntex' expense. The procedure for sample preparation and analysis of fish fillets is set forth in Attachment B. Splits of all samples will be prepared and maintained by the laboratory and will be provided by Syntex to EPA, upon request.

B. Sediment. Within the period August 1 to August 31 of each year during the five year initial period of sampling, sediment samples will be obtained by EPA (or its designee) at

Locations 1, 3 and 5. The quantity of each sediment sample will be at least 2 kg. The samples shall be stored in glass containers with teflon or aluminum foil lid liners. EPA has provided Syntex a brief description of the methods for sediment collection, split preparation and handling, a copy of which may be found at Attachment F. Such methods shall be consistent, in material respects, from year to year. The completed Sampling Log will be prepared by EPA or its designee for each sediment sample. Such Sampling Log will be sent by EPA or its designee to Syntex. Splits of all sediment samples will be prepared by EPA or its designee and provided to Syntex. Syntex will take custody of the sample splits for analysis after they have been collected and labeled by EPA or its designee. Analysis of such samples shall be performed by Syntex or its designee.

III. ANALYTICAL PROCEDURES

A. Fish. Attachment B describes the procedure to be used for isomer-specific analysis of TCDD in fish fillets. The sensitivity of the analytical procedure will be reported for each sample. It is anticipated that this sensitivity level will be approximately 5-15 parts per trillion ("ppt"). A capillary column Gas Chromatograph Mass Spectrometer ("GCMS") will be used in all fish analyses.

B. Sediment. The procedure for analysis of TCDD in soil and sediment is attached as Attachment C. The sensitivity

of the analytical procedure will be reported for each sample. It is anticipated that the sensitivity level will be approximately 10 parts per trillion ("ppt"). A capillary column GCMS will be used in all sediment analyses.

IV. EXTENSION OF MONITORING BEYOND FIVE YEARS

Any extensions of fish or sediment sampling will be governed by the provisions of paragraph 42 of the Consent Agreement and Order.

V. STATISTICAL ANALYSIS

A. Fish. The data to be statistically analyzed will consist, initially, of two independent TCDD measurements ("Data Points") at each of five locations at each of five time points one year apart. Each Data Point will be the result of an analysis of a homogenate of the fillets from four to ten fish. If less than eight fish are obtained from any location, a single homogenate will be prepared and analyzed. A value of one-half the detection limit of the assay will be assigned to all samples which fall below the detection limit. Any analysis having a detection limit above 15 ppt will be repeated, if practicable, or removed from the statistical analysis of the data.

The ten Data Points from Location 1 (0.3 miles downstream) will be statistically evaluated separately from those collected at the other downstream locations. The Jonckheere test, a

nonparametric test for ordered alternatives, will be used to test:

H_o: $T1 = T2 = T3 = T4 = T5$ (TCDD level unchanged over 5 years)

versus

H_a: $T1 \geq T2 \geq T3 \geq T4 \geq T5$ where at least one inequality is strict.

This test is used to detect a monotonic decrease in the TCDD level over time. In addition, a logarithmic transformation will be applied to the data to normalize the distribution and stabilize the variance. The resulting values will be analyzed by means of least squares linear regression using the model

$$\ln Y(i) = \ln B0 + (B1 \times T) + e(i)$$

where Y(i) is the (i)th measurement of TCDD concentration, B0 and B1 are constants, T is the year in which the measurement was taken, and e(i) is a random error term. A one-tailed t-test will be carried out at a significance level, alpha, of 0.05 to test

$$H_o: B1 = 0$$

versus

$$H_a: B1 < 0$$

in order to determine whether the slope of the fitted line is decreasing. Graphs will be drawn of the raw data versus time, the transformed data versus time, and the residuals

(differences between the actual data points and those predicted by the linear regression) versus time. Selection of whether to rely upon parametric or non-parametric analyses will be based on the method which has the highest level of significance.

The data from the remaining four downstream (3, 6, 9 and 12 mi.) locations will be analyzed similarly. The Jonckheere test will be applied to data from each location separately and the p values will be combined using Fisher's method (Fisher, R. A., Statistical Methods for Research Workers, Oliver and Boyd, 1958, pp. 99-101). Multiple linear regression will be carried out on the natural log transform of the forty data points using the model

$$\ln Y(i) = \ln B0 + (B1 \times T) + (B2 \times X) + e(i)$$

where Yi is the (i)th TCDD measurement, B0, B1, and B2 are constants, T is the year, X is the distance to the location, and e(i) is a random error term. A confidence interval for B1 will be constructed, since the goal in this case is to show whether or not the TCDD levels are increasing. The significance level used will be alpha=0.10 in order to decrease the probability of a Type II error. The raw data, the transformed data, and the residuals will be graphed versus time.

B. Sediment. The data to be analyzed will consist of one measurement at each of three locations at each of five sampling times. The multiple linear regression technique

outlined for the fish samples will be applied to the logarithmic transform of the sediment sample values.

C. Statistical Methods.

1. Jonckheere test

(Ref. Hollander, M. and Wolfe, D., Nonparametric Statistical Methods, John Wiley and Sons, 1973, pp. 120-123).

Ho:T1 = T2 = T3 = T4 = T5

Ha:T1 $\geq$ T2 $\geq$ T3 $\geq$ T4 $\geq$ T5 where at least one inequality is strict.

For each pair of sampling times, i and j, calculate

$$U(i,j) = \sum_{\mu=1}^{2} \sum_{\nu=1}^{2} \emptyset(X(u,i), X(v,j))$$

where u is the sample size in year i and v is the sample size in year j and

$$\begin{aligned} \bigcirc(a,b) &= 1 \text{ if } a > b \\ &= 1/2 \text{ if } a = b \\ &= 0 \text{ if } a < b \end{aligned}$$

U(i,j) is a count of the number of times a data point from year (i) is larger than a data point from year (j). Since there are ten possible pairs of sampling times, (five years taken two at a time) there are ten values of

U(i,j). Let J = $\sum_{i<j}$ U(i,j) and compare J to Table A.8 in Hollander and Wolfe.

2. Linear regression

(Ref: Draper, N.L., and Smith, H., Applied Regression Analysis, John Wiley and Sons, 1966, pp. 7-20).

Using the model:

$$\ln Y(i) = \ln B0 + (B1 + T) + e(i)$$

calculate the estimates of the slope

$$b(1) = \frac{n\sum t \cdot \ln y - (\sum t)(\sum \ln y)}{n\sum t^2 - (\sum t)^2}$$

and intercept

$$\ln b(0) = \overline{\ln y} - b(1) \times \overline{t}$$

To test the hypothesis that the slope is negative, calculate

$$t = \frac{b\sqrt{\sum t_2 - (\sum t)_2/n}}{s}$$

where $s = \sqrt{\text{residual sum of squares}/(n-2)}$

and compare t to Student's t for alpha=0.05 and n-2 degrees of freedom.

3. Multiple linear regression

(Ref: Draper, N.R., and Smith, H., <u>Applied Regression Analysis</u>, John Wiley and Sons, 1966, pp. 65, 104-124).

Using the model

$$\ln Y(i) = \ln B0 + (B1 \times T) + (B2 \times X) + e(i)$$

calculate the estimates ln b0, b1, and b2 using the matrix approach

$$b = (X'X)^{-1} X'Y$$

A 90% confidence interval for the coefficient B1 is calculated as follows:

$$CI = b1 + t(n-p-1, 0.95)\sqrt{C(11)} \times s$$

where C(11) is the diagonal element of the matrix (X'X) corresponding to T, p is the number of independent variables and s is the square root of the residual sum of squares divided by n-p-1 degrees of freedom.

VI. REPORTS

Annual reports containing the final results of the Spring River fish fillet and sediment analyses will be provided promptly to EPA within 15 days of their completion and acceptance by Syntex, after verification that the proper procedures and calculations have been performed. Such annual report will include the relevant graphs and statistical computations. A final report shall be supplied to the Regional Administrator, Region VII, EPA at the end of the five year sampling program in accordance with the requirements of paragraph 47 of the Consent Agreement and Order. Such final report will include the relevant graphs and statistical computations.

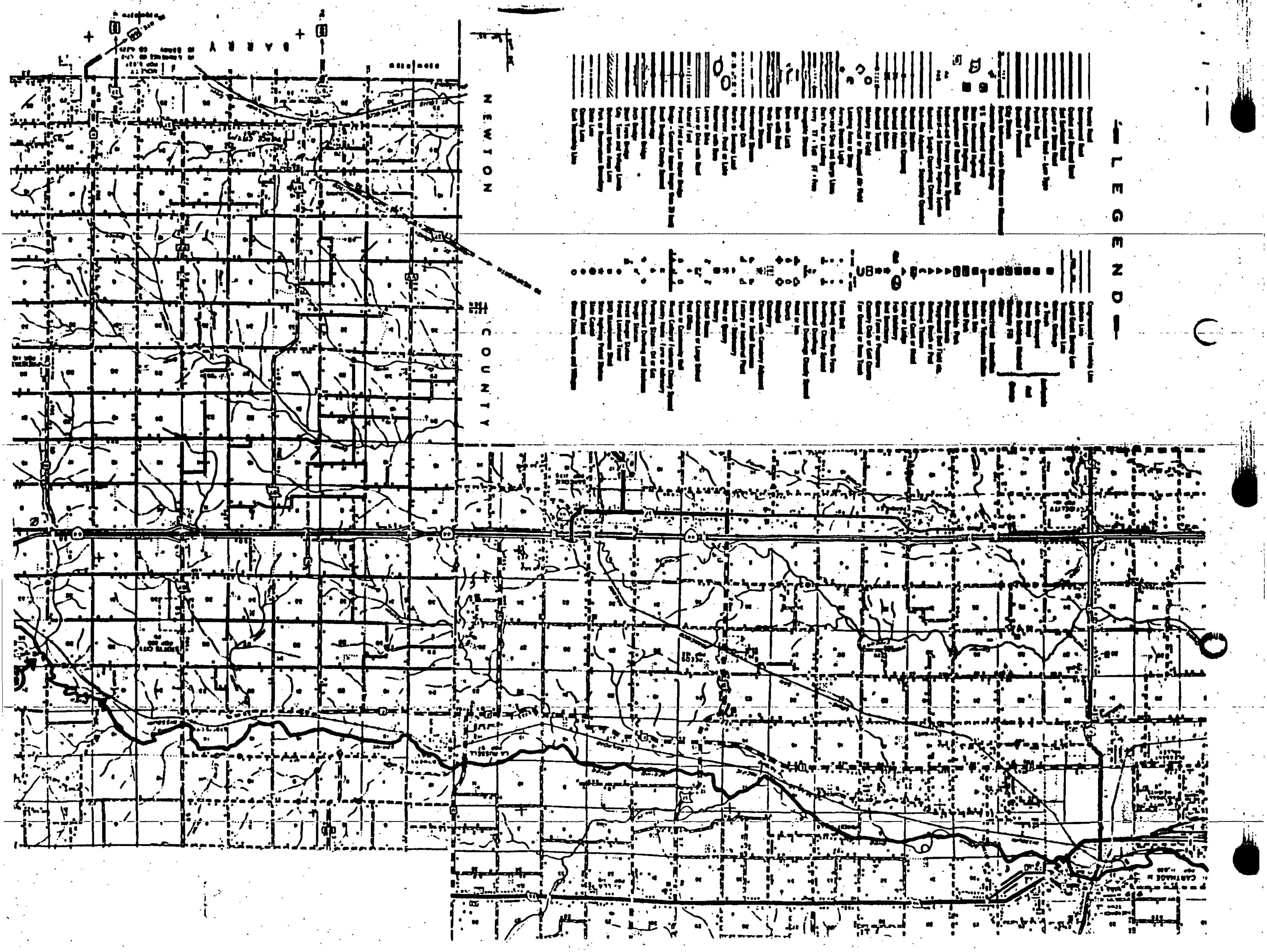
VII. SCHEDULE

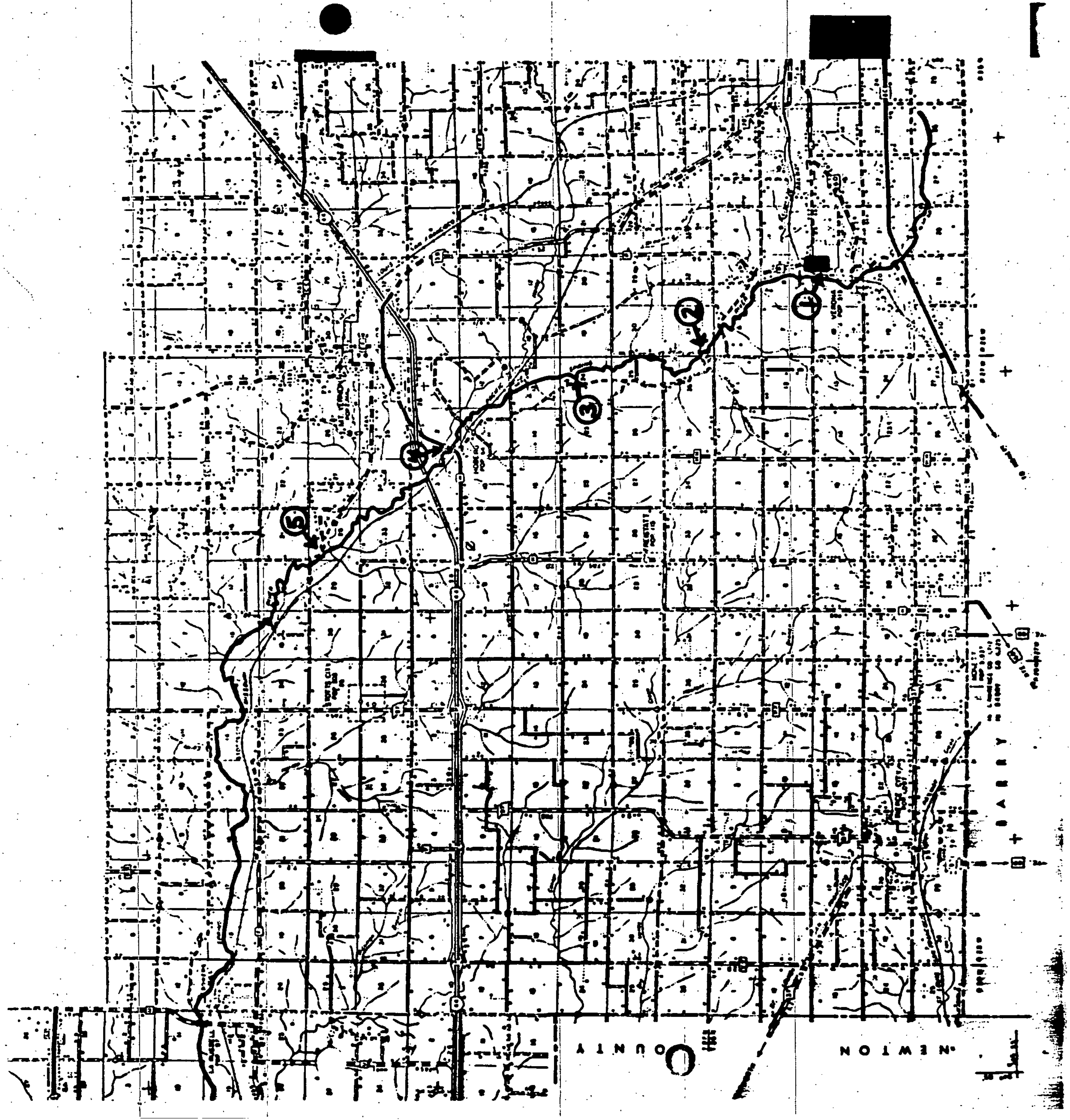
A. Dr. Gross (or such other laboratory as may be selected) will be requested to complete the analyses of fish fillet composites and sediment samples within 45 days of receipt of the samples from EPA.

B. Annual reports of analyses will be completed within 15 days of receipt and acceptance by Syntex of all analytical data.

C. The final five-year report will be supplied within 30 days of receipt and acceptance by Syntex of all analytical data.

Attachment A


BARRY
NEWTON COUNTY
LEGEND
CARTHAGE


1
2
3
4
5
BARRY
COUNTY
NEWTON

Description of Fish Sampling Locations

Location 1 - 0.3 miles at junction of Spring River and "Farm Road" bridge. 1100 feet west of intersection of State Highway P and "Farm Road."

Location 2 - Three miles along "Farm Road" approximately 3500 feet south of Spring River Chruch. Airphoto 4

Location 3 - Six miles along "Farm Road" approximately two miles south of Hoberg, Missouri, Airphoto 6.

Location 4 - Nine miles at junction of Spring River and State Highway H bridge. Airphoto 9.

Location 5 - Twelve miles at junction of Spring River and State Highway V. Air photo 11.

United States
Environmental Protection
Agency

Environmental Monitoring
Systems Laboratory
P.O. Box 15027
Las Vegas, NV 89114

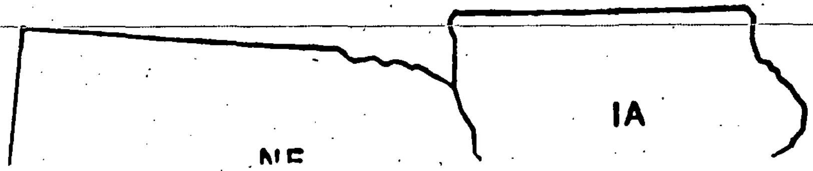
TS-AMD-83059
December 1983

Research and Development

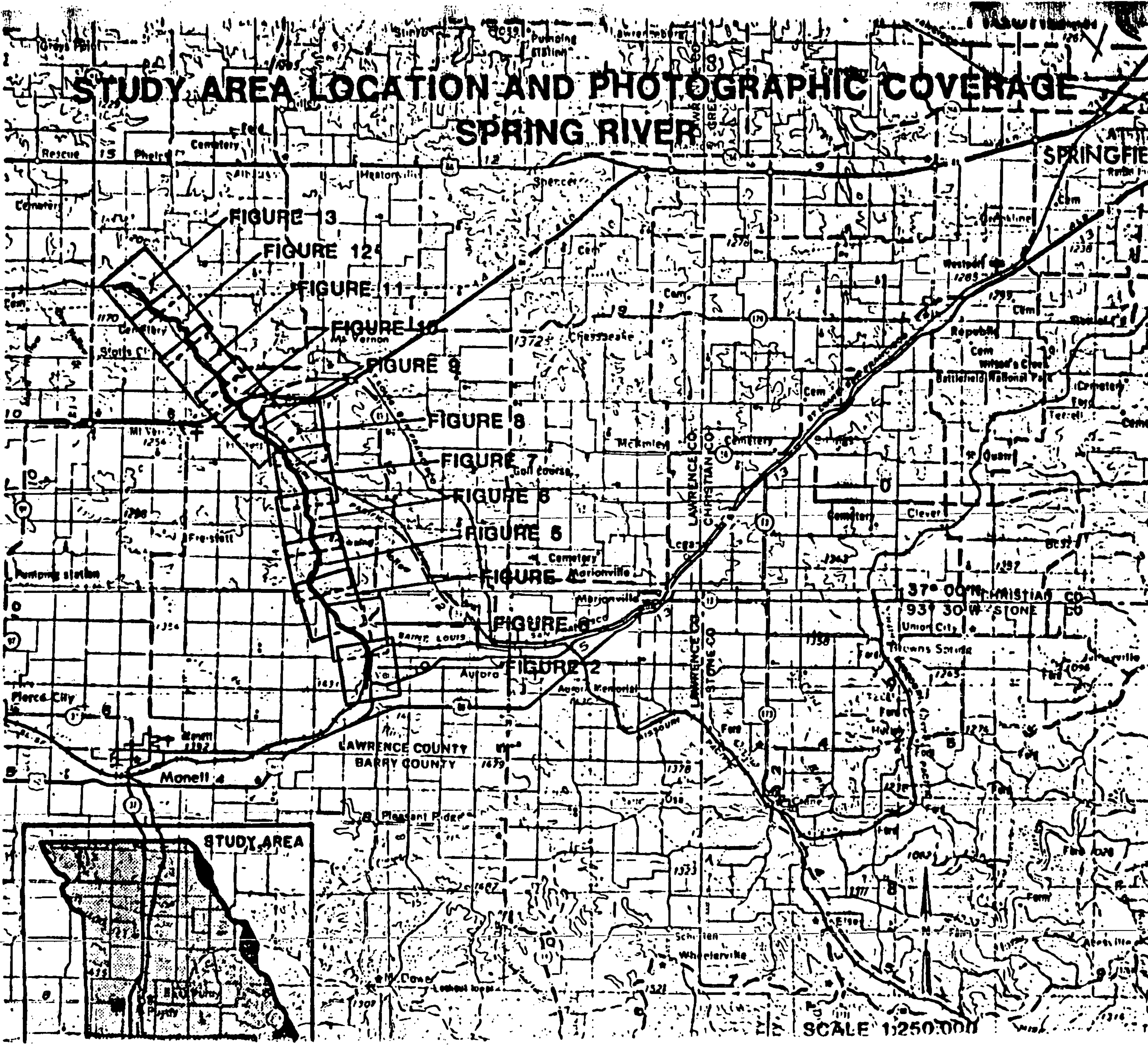
EPA

AERIAL PHOTOGRAPHIC ANALYSIS OF LAND USE ACTIVITIES
Spring River, Missouri
May 1983

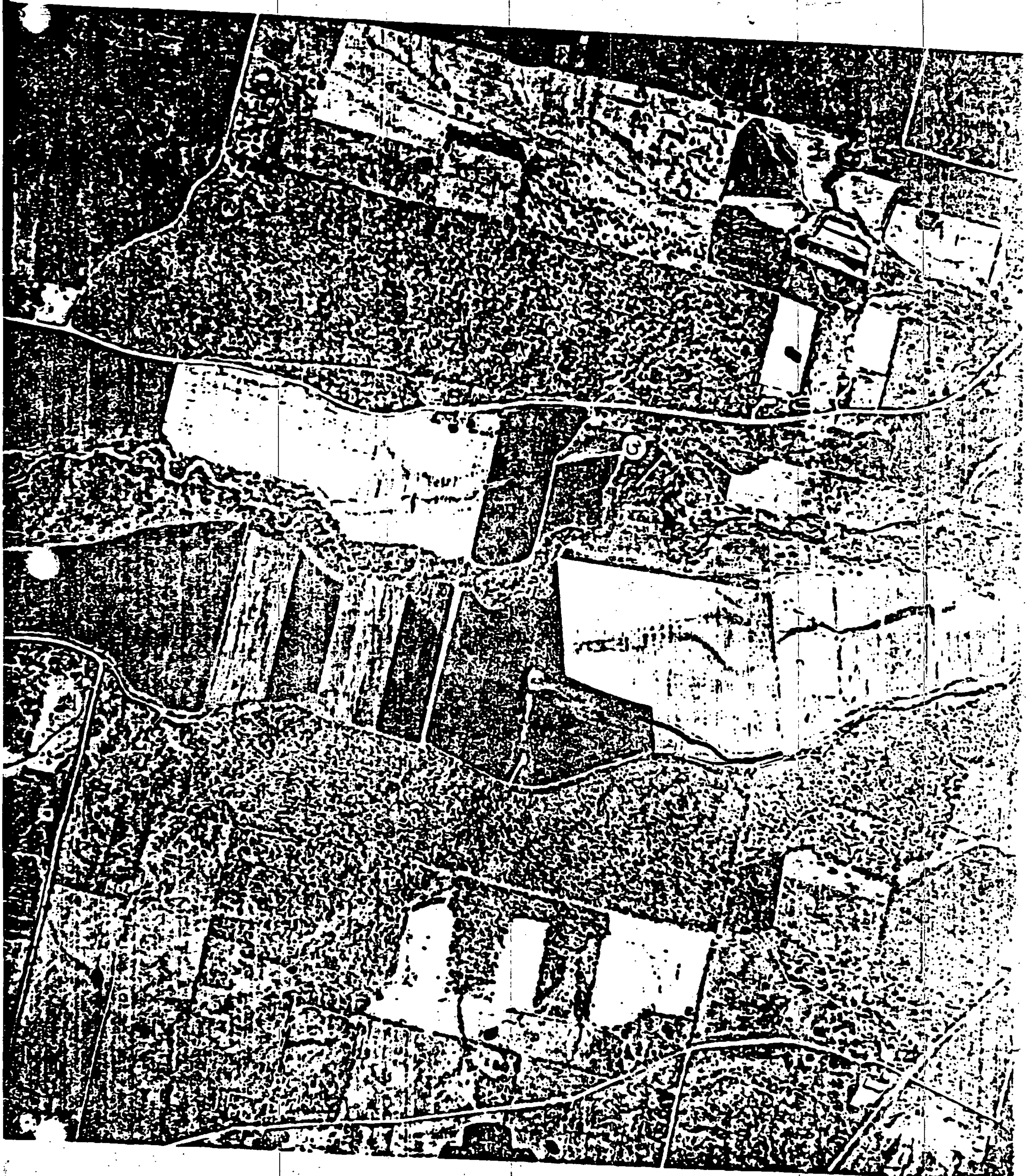
EPA Region 7



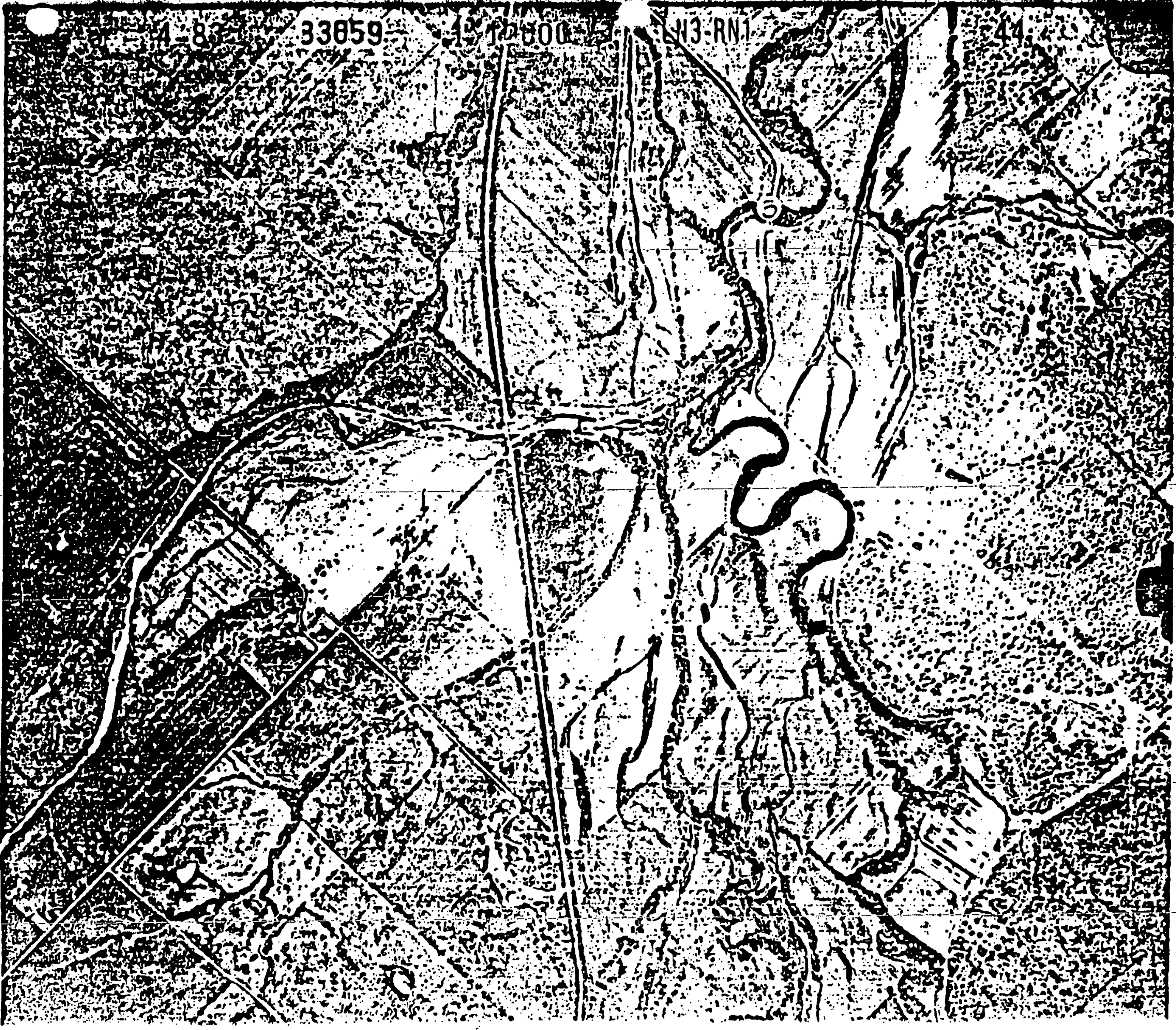
STUDY AREA LOCATION AND PHOTOGRAPHIC COVERAGE
SPRING RIVER



2


G

330591
1:12000
(3801)
48
4


4-83
33059
1:12000
N3-RN1
44

Attachment B

8403.02
8351G(8311G)/475

Attachment B

Isomer Specific Analysis of 2,3,7,8-Tetrachlorodibenzodioxin (TCDD) in Fish Fillets

SAMPLE PREPARATION

Packages of ten frozen fish from each sampling point will be placed in a freezer immediately upon delivery. All fish will remain in the frozen state until the time of analysis. The fish from each group of ten will be thawed and filleted. The fillets will be cut into smaller pieces and homogenized. All fillets will be used to make up the respective homogenates. Two approximately equal portions of the homogenized samples of each group will be collected in glass jars, covered with aluminium foil-lined lids and labeled. One of the jars will be frozen and preserved for any split samples that EPA may request. In year-1, the fish homogenate in the second jar will be divided into five approximately equal portions. Four of the portions will be placed in similar containers, labeled and frozen for analysis in subsequent years. The remaining portion will be utilized for the current analysis.

SAMPLE EXTRACTION PROCEDURE FOR TISSUE

A 5-10 g sample is accurately weighed and spiked with a known amount (2.0-2.5 ng) of $^{13}C_{12}$-TCDD. It is then saponified in 15 ml of ethanol and 30 ml of 40% aqueous KOH in a reflux apparatus until completely hydrolyzed.

The solution is transferred to a 250 ml separatory funnel and diluted with 20 ml of ethanol and 40 ml of water and extracted four time with nanograde hexane. The first extraction is done with 25 ml of hexane, shaking vigorously for one minute. The lower aqueous layer is removed to a clean beaker, and the upper hexane layer decanted to a 125 ml separatory funnel. The aqueous layer is then extracted three times more with 15 ml portions of hexane, each time adding the hexane to the 125 ml separatory funnel. The combined hexane extracts are washed with 10 ml of water to remove excess base.

The combined hexane extracts are washed 4 times wih 10 ml concentracted H_2SO_4, or until both layers are clear. As many as 8 extracts may be necessary, depending on the sample. Again the hexane is washed with 10 ml of water. The hexane layer is decanted to a 2 ounce jar and concentrated under a stream of dry nitrogen to approximately 1 ml.

LIQUID CHROMATOGRAPHY CLEAN-UP

Silica Chromataography

A 5 cm column is prepared using a disposable pipet plugged with glass wool. The silica is capped with 1/4 cm anhydrous sodium sulfate and then wetted with hexane. The sample, dissolved in one ml of hexane, is transferred to the column. TCDD is eluted with 3 ml of 20% (V/V) benzene in hexane. All the eluate is collected and concentrated to one ml. Additional hexane is added, and the sample was again evaporated to one ml to reduce the proportion of benzene.

Alumina Chromatography

The alumina is prepared by saturating with methylene chloride, removing excess solvent, then activating at 165°C for 24 hours. A column is prepared in the same manner as the silica column above. The column is cooled to room temperature in a dessicator before use.

Hexane is used to wet the column before transferring the sample. The alumina is eluted with 6 ml of pesticide grade CCl_4, then with 4 ml of 10% CH_2Cl_2 and finally with 6 ml of CH_2Cl_2. The methylene chloride/hexane fraction is collected and concentrated under nitrogen while replacing the volatile CH_2Cl_2 with hexane. All other fractions are discarded.

LIST OF MATERIALS USED IN SAMPLE EXTRACTION

Acetone, OmniSolv, MCB

Benzene, OmniSolv, MCB

Carbon tetrachloride, OmniSolv, MCB

Ethyl alcohol, OmniSolv, MCB

Hexane, OmniSolv, MCB, non U:V

Methylene chloride, OmniSolv, MCB

Sulfuric acid, concentrated, analytical reagent, Mallinckrodt

Water, distilled in glass

Potassium hydroxide, analytical grade, Mallinckrodt

Sodium sulfate (anhydrous), analytical grade, Fisher

Sodium carbonate (anhydrous), analytical grade, Fisher

Aluminum oxide, neutral, activity grade I, Woelm Pharma

Silica gel, 60-200 mesh, reagent grade, Baker Chemical Co.

Dry nitrogen (boil-off from liquid N_2)

All OmniSolv line solvents are distilled in glass, suitable for chromatography and residue analysis.

LIST OF MATERIALS USED IN SAMPLE EXTRACTION

Acetone, OmniSolv, MCB

Benzene, OmniSolv, MCB

Carbon tetrachloride, OmniSolv, MCB

Ethyl alcohol, OmniSolv, MCB

Hexane, OmniSolv, MCB, non U:V

Methylene chloride, OmniSolv, MCB

Sulfuric acid, concentrated, analytical reagent, Mallinckrodt

Water, distilled in glass

Potassium hydroxide, analytical grade, Mallinckrodt

Sodium sulfate (anhydrous), analytical grade, Fisher

Sodium carbonate (anhydrous), analytical grade, Fisher

Aluminum oxide, neutral, activity grade I, Woelm Pharma

Silica gel, 60-200 mesh, reagent grade, Baker Chemical Co.

Dry nitrogen (boil-off from liquid N_2)

All OmniSolv line solvents are distilled in glass, suitable for chromatography and residue analysis.

2,3,7,8-Isomer Specific TCDD Analysis by Capillary Column GC/HRMS

Appropriate dilutions of the samples will be made with hexane at the time of analysis and the aliquots from the resulting solutions will be used for capillary column GC/HRMS.

A. Gas Chromatography/Mass Spectrometer

A Kratos MS-80 medium resolution mass spectrometer (ultimate resolution 20,000), equipped with a 5 channel multiple peak monitoring (MPM) device will be used. The mass spectrometer is coupled to a Carlo-Erba Gas chromatograph equipped with a SE-54 fused silica capillary column (0.25 mm X 30 m).

B. Gas Chromatographic Conditions

Typical operating conditions: Helium with a linear velocity of 35 cm/sec, injector 250°C, detector 275°C, column temperature 150°C, isothermal for 10 minutes, and then programmed at 5°/min to 280°C. The sample is injected in hexane at a temperature of 57°C. The split/sweep valves are closed for 2 minutes after injection.

LIST OF MATERIALS USED IN SAMPLE EXTRACTION

Acetone, OmniSolv, MCB

Benzene, OmniSolv, MCB

Carbon tetrachloride, OmniSolv, MCB

Ethyl alcohol, OmniSolv, MCB

Hexane, OmniSolv, MCB, non U:V

Methylene chloride, OmniSolv, MCB

Sulfuric acid, concentrated, analytical reagent, Mallinckrodt

Water, distilled in glass

Potassium hydroxide, analytical grade, Mallinckrodt

Sodium sulfate (anhydrous), analytical grade, Fisher

Sodium carbonate (anhydrous), analytical grade, Fisher

Aluminum oxide, neutral, activity grade I, Woelm Pharma

Silica gel, 60-200 mesh, reagent grade, Baker Chemical Co.

Dry nitrogen (boil-off from liquid N_2)

All OmniSolv line solvents are distilled in glass, suitable for chromatography and residue analysis.

C. Mass Spectrometric Conditions and Multiple Ion Selection

The mass spectrometer is operated in the EI mode (70eV, 250°C) at 7500 resolving power. Peak profiles are acquired at an amplified bandwidth of 30 KHz. The ions m/z 319.8965, m/z 321.8936 and m/z 333.9339 (^{13}C-2,3,7,8-TCDD) are monitored. The instrument is tuned using m/z 330.9792 of PFK, and this ion is used as a check mass on channel 4. The output of the mass spectrometer is recorded on a 3-pen strip chart recorder (Linear Model-595).

D. Calculation of Results

Quantification is achieved using the internal standard ratio method. Throughout the experiment, standard samples containing 2,3,7,8-TCDD and ^{13}C-2,3,7,8-TCDD are analyzed. The slopes of the calibration plots are computed as the averages of the ratios of $(I^{334}/ng)/(I^{332}/ng)$ (I is the normalized intensity for the designated mass).

Residue levels of TCDD in actual samples are calculated by comparing the ratios of intensities of I^{322}/I^{334} obtained for a given sample with the slope of the calibration plot. The detection limit is considered to be the respective value obtained for an intensity of 2.5 x noise level measured at the base line.

The internal standard (^{13}C-2,3,7,8-TCDD) is utilized in the calculation of percent recoveries, and in doing so the abolute intensity (I^{334} normalized) is measured and compared with the intensities (I^{334}/ng) obtained by injecting standard solutions of the internal standard.

Validation

Results which indicate the presence of TCDD will be validated by comparing the signal intensities of m/z 319.8965 and m/z 321.8936 (the two most abundant ions of TCDD). The theoretical ratio of m/z 320/322 is 0.77. Validation of TCDD is considered acceptable if the observed ratio of signals is 0.77 ± 0.10.

The retention times of the isomers are measured from the point of injection and normalized to the position of the signal of the internal standard, ^{13}C-2,3,7,8-TCDD.

Attachment C

SYNTEX RESEARCH METHOD NO. 10,317A

DETERMINATION OF 2,3,7,8-TETRACHLORODIBENZO-P-DIOXIN (2,3,7,8-TCDD) IN SOIL, SEDIMENT AND SLUDGE BY CAPILLARY GAS CHROMATOGRAPHY LOW RESOLUTION MASS SPECTROMETRY SELECTED ION MONITORING (C-GC-LRMS-SIM)

SCOPE AND APPLICATION

This method is intended for use in the determination of 2,3,7,8-TCDD in soil, sediment and sludge. The specificity of the method for the 2,3,7,8-TCDD isomer is dependent upon the capillary column used rather than the mass spectrometer. An isomer test mixture must therefore be injected to determine isomer specificity for any column used.

The linear range of the analysis depends upon three variables, the amount of sample extracted, the amount of internal standard added, and the amount of interference from background contamination in the sample matrix. For a 200 g sample with a 10 ng internal standard spike the expected range is from 0.003 ppb to 2.5 ppb. For a 20 g sample with 10 ng or 100 ng of internal standard the expected range is from 0.03 ppb to 25 ppb or 0.3 ppb to 250 ppb respectively. For a 5 g sample with 100 ng of internal standard the expected range is 1.0 ppb to 1000 ppb. In addition detection limits below 0.1 ppb can be achieved only when the amount of interference from background contamination in the sample is minimal.

The method is recommended for use only by the experienced analyst, or by technicians well briefed and under the supervision of an analyst trained in the handling of TCDD. Because of the reported toxicity of 2,3,7,8-TCDD precautions must be taken to prevent exposure to personnel by materials known or believed to contain 2,3,7,8-TCDD.

SUMMARY

A nominal sample aliquote of 20 g is used for preparation and analysis. The wet sample is weighed, combined with sodium sulfate and dried. It is spiked with 10 ng or 100 ng of $^{13}C_{12}$-2,3,7,8-TCDD internal standard, depending on the analytical linear range required, then soxhlet extracted using dichloromethane. The extract is concentrated and filtered through activated silica using 10/90 dichloromethane/hexane. The filtrate is then exchanged into hexane for chromatography cleanup with two columns containing modified silica and basic alumina. The final isolated sample is brought to a final volume of 50 µl or 200 µl corresponding to the internal standard spike of 10 ng or 100 ng respectively. The samples are then quantitated for 2,3,7,8-TCDD by Capillary-Gas Chromatography-Low Resolution Mass Spectromety-Selected Ion Monitoring (C-GC-LRMS-SIM).

APPARATUS AND MATERIALS

1. Soxhlet extractors - 40 mm ID. with paper thimbles, 35 x 90 mm, and 250 ml boiling flasks. Soxhlet Extractors - 65 mm ID, with paper thimbles, 60 x 180 mm, and 1000 ml boiling flasks.
2. Chromatography columns - 0.8 cm ID. x 20 cm and 1.5 cm ID. x 30 cm equipped with coarse glass frits and stopcocks.
3. Separatory funnels - 250 ml.
4. Round bottom flasks - 250 ml.
5. Fritted glass filters - coarse, 60 ml, 30 ml.
6. Volumetric flasks - 1 ml, 10 ml, 25 ml, 100 ml.
7. Volumetric pipettes - 1 ml, 5 ml.
8. Microliter pipettes - 50 lambda, 100 lambda, 200 lambda.
9. Disposable transfer pipettes - Pasteur type with bulbs.
10. Glass vials - 20 ml (silanized).
11. Cone shaped vials - 1 ml (silanized).
12. Alundum boiling stones - Soxhlet extracted using toluene, vacuum dried at 150°C for 20 hours.
13. Rotary vacuum evaporator.
14. Ultrasonic water bath.
15. Top loading electronic balance capable of accurately weighing 20 g to the nearest 0.01 g.
16. Capillary Gas Chromatograph coupled to a Low Resolution Mass Spectrometer equipped for splitless injection. HP 5790A/5970A or equivalent. Direct interface recommended.
17. Recommended capillary column - 25 m x 0.20 mm, fused silica, 0.33 μ film crosslinked 5% phenyl methyl silicone, available through HP.

NOTES: A. All glassware is initially cleaned with aqueous detergent then rinsed with water, methanol, dichloromethane and hexane. Thereafter, between samples, glassware is rinsed with the series of solvents.

B. Silanized glassware is prepared by a 5 minute treatment with 5% dichlorodimethylsilane in toluene at room temperature, followed by a rinse with methanol.

REAGENTS

1. Hexane - Pesticide quality distilled in glass or equivalent.
2. Toluene - Pesticide quality distilled in glass or equivalent.
3. Dichloromethane - Pesticide quality distilled in glass or equivalent. NOTE: The dichloromethane and toluene may need to be redistilled in the laboratory using a glass distillation column.
4. Methanol - A.R.
5. Acetone - A.R.
6. Water - Deionized, purified through activated carbon, and one liter batches extracted three times with 400 ml hexane prior to use.
7. H_2SO_4 - concentrated A.R.
8. NaOH - A.R.
9. $AgNO_3$ - A.R.
10. n-Tetradecane - Applied Science Supplier, or equivalent.
11. Na_2SO_4 - anhydrous, granular, Soxhlet extracted using dichloromethane/hexane 40/60 (^{v}/v), vacuum dried at 150°C for 20 hours.

REAGENTS (Con't.)

12. Alumina – E. Merck Basic (70/230 Mesh) or equivalent, Soxhlet extracted using dichloromethane/hexane 40/60 (^{v}/v), rinsed with toluene followed by hexane, vacuum dried and activated at 150°C for 20 hours.
The activated alumina should be protected from moisture in storage.

13. Silica –Silica Gel 60 70/230 mesh, E. Merck or equivalent, Soxhlet extracted using dichloromethane/hexane 40/60 (^{v}/v), rinsed with hexane, vacuum dried and activated at 150°C for 20 hours.

14. 44% H_2SO_4 modified Silica, 100 g – 24 ml of concentrated H_2SO_4 is combined with 56.0 g of activated silica gel. The mixture is shaken in a closed container until it is homogeneous and free-flowing.

15. 33% Aqueous 1 M NaOH modified Silica, 100 g – 100 ml of 1 M NaOH is extracted three times with 50 ml hexane. 32 ml of the extracted 1 M NaOH is combined with 67.0 g of activated silica gel. The mixture is shaken in a closed container until it is homogeneous and free-flowing.

16. 10% $AgNO_3$ modified silica, 100 g – 10.0 g of $AgNO_3$ is dissolved in 25 ml of water. The aqueous $AgNO_3$ is extracted three times with 30 ml hexane then combined with 90.0 g of activated silica gel. The mixture is shaken in a closed container until it is homogeneous and free-flowing. After allowing the mixture to stand for 2 hours it is vacuum dried and activated at 150°C for 20 hours. The 10% $AgNO_3$ modified silica should be protected from light.

17. 2,3,7,8–TCDD Stock Standard Solutions

A. 50 µg/ml solution in isooctane which can be purchased from commercial sources (Cambridge Isotope Laboratories, Inc.), 141 Magazine St., Cambridge MS. 02139.

B. 5.0 µg/ml – Prepared by a 1.0 ml to 10.0 ml dilution of solution A into toluene.

C. 500 ng/ml – Prepared by a 1.0 ml to 10.0 ml dilution of solution B into toluene.

D. 50 ng/ml – Prepared by a 1.0 ml to 10.0 ml dilution of solution C into toluene.

E. 5.0 ng/ml – Prepared by a 1.0 ml to 10.0 ml dilution of solution D into toluene.

Note: The 500 ng/ml standard must be verified by analysis against a certified 2,3,7,8–TCDD standard. A certified 2,3,7,8–TCDD check solution is available from EPA (Environmental Monitoring Systems Labs – Las Vegas) at a concentration of 7.87 µg/ml in isooctane.

18. $^{13}C_{12}$–2,3,7,8-TCDD Stock Spiking Solutions

A. 50 µg/ml solution in isooctane which can be purchased from commercial sources (Cambridge Isotope Laboratories).

B. 1.0 µg/ml – Prepared by a 1.0 ml to 50.0 ml dilution of solution A into 10/90 toluene/acetone.

C. 100 ng/ml – Prepared by a 1.0 ml to 10.0 ml dilution of solution B into 10/90 toluene/acetone.

Note: The 1.0 µg/ml solution must be demonstrated to contain a relative abundance for ions 320 and 322 less than 1% of ions 332 and 334 respectively.

19. Column Performance Solution
A toluene solution containing approximately 100 ng/ml each of seven TCDD isomers (2378, 1478, 1234, 1237, 1238, 1278, and 1267) and $^{13}C_{12}$-2,3,7,8-TCDD.

The seven isomer mixture is available from EPA (Environmental Monitoring Systems Labs - Las Vegas).

SAMPLE PREPARATION

Extraction of Soil, Sediment or Sludge

Prior to each extraction, thimbles are placed into their extractors and fresh dichloromethane is refluxed through the system at a recycle rate of 20 ml per minute for 2 hour. The thimbles are removed and vacuum dried, and the extractors are rinsed with hexane.

A representative 20 g portion of sample is weighed, mixed with 30 g sodium sulfate in a beaker and allowed to air dry in a fume hood. The sample is periodically remixed during the first one or two hours of drying to prevent the formation of clumps. Drying is then continued overnight. The dried sample is broken up and remixed then transferred to a preextracted paper thimble. The sample is spiked with 100 ng of $^{13}C_{12}$-2,3,7,8-TCDD (100 µl of the 1.0 µg/ml spiking solution) and covered with a layer of sodium sulfate. It is placed into the extractor and extracted at a reflux rate of 20 ml/min. for 6 hours using 130 ml of dichloromethane. One drop of n-Tetradecane is added to the extract and the sample is evaporated to dryness using a rotary vacuum evaporator.

The sample extract is redissolved in 50 ml of 10/90 dichloromethane/hexane and filtered through 4 g of activated silica in a 30 ml scintered glass Buchner funnel. The flask is then rinsed twice with 15 ml of 10/90 dichloromethane/hexane through the filter. One drop of n-tetradecane is added and the sample is evaporated to dryness. It is then redissolved in 5 ml of hexane for column chromatography. If the residue is not soluble in 5 ml of hexane the filtration procedure is repeated using 10/90 dichloromethane/hexane.

Note: The sample size and quantity of $^{13}C_{12}$-2,3,7,8-TCDD spike are adjusted to fit the expected range of the analysis.

Column Chromatography Clean-up

Two columns are prepared. Column A (1.5 x 30 cm) contains from bottom to top, 1.5 g actived 10% silver nitrate on silica gel, 1.0 g actived silica gel, 2.0 g 33% 1N sodium hydroxide on silica gel, 1.0 g activated silica gel, 5.0 g 44% concentrated sulfuric acid on silica gel, 2.0 g activated silica gel, and 3.0 g sodium sulfate. Column B (0.8 x 20 cm) contains from bottom to top 3.0 g of activated basic alumina, and 2.0 g sodium sulfate.

Both columns are wetted with hexane and Column A is positioned above column B such that all effluent from Column A passes through Column B. The sample in 5 ml hexane is applied and the flask rinsed twice with 5 ml hexane onto Column A. The hexane is allowed to elute to the top of the sodium sulfate between rinses. Column A is then eluted with 50 ml of hexane and is discarded. The hexane is eluted to the top of the sodium sulfate in Column B and Column B is eluted with 10 ml of 15/85 dichloromethane/hexane followed with 15 ml of 30/70 dichloromethane/hexane.[1] The later 15 ml fraction is collected in a silanized vial containing 5 µl of n-tetradecane and is evaporated to dryness using a filtered nitrogen stream. The sample is redissolved in a minimum of toluene and transfered quantitatively to a 1 ml silylized cone shaped vial then evaporated to a final volume of about 200 µl for quantitation by C-GC-LRMS-SIM.

[1] Note: The proper amount of 15% (^{v}/v) dichloromethane in hexane and 30% (^{v}/v) dichloromethane in hexane is highly dependent upon the activity of the alumina. Experience has shown that for each batch of alumina, the volumes of 15% and 30% (^{v}/v) dichloromethane in hexane should be re-determined. A test column chromatography run is performed using an aliquot 2,3,7,8-TCDD standard in hexane. By collecting 4.0 ml fractions of the 15% solution the analyst must first determine what volume of 15% solution can be eluted before any 2,3,7,8-TCDD can be detected in the fractions by C-EC-GC. Then by eluting with the 30% solution the analyst must determine what volume of 30% solution must be eluted to recover at least 95% of the added 2,3,7,8-TCDD.

QUANTITATION BY C-CG-LRMS-SIM

C-GC-MS-SIM Analysis Conditions

Analyses are performed using an HP 5790A Capillary Gas Chromatograph coupled to an HP 5970A Low Resolution Mass Spectrometer using a direct interface.

Ions m/z 321.9, 319.9 resulting from native 2,3,7,8-TCDD and m/z 334.0, 332.0 resulting from the $^{13}C_{12}C$-2,3,7,8-TCDD internal standard are monitored under the following conditions:

Column:	25 m x 0.20 mm, 0.33 µ film crosslinked 5% phenyl methyl silicone, fused silica capillary column, HP supplier.
Carrier:	He, 20 psi head pressure
Injector:	250°C, splitless 0.80 min. delay
Inj. Vol.:	1.6 µl
Col. Temp.:	130°C initially for 2 minutes Programed 20°C/min. to 220°C Programed 8°C/min. to 280°C Programed 20°C/min. to 300°C Hold at 300°C for 7 minutes
Mass Detector:	Standard autotune conditions, 0.2 AMU window Dwell Times 60 msec each ion.
2,3,7,8-TCDD Retention Time:	12.2 minutes

Calibration

A. Verification of 2,3,7,8-TCDD Stock Standard Solutions:

A 1.0 ml to 10.0 ml dilution of the 7.87 µg/ml 2,3,7,8-TCDD check solution is made into toluene. This results in a 787 ng/ml solution.

Four vials are then prepared in silanized 1 ml cone shaped vials. Into each vial 5 µl of n-tetradecane and 100 µl of the 1.0 µg/ml $^{13}C_{12}$-2,3,7,8-TCDD spiking solution are added. Into two vials 100µl of the 500 ng/ml 2,3,7,8-TCDD stock standard solution is added. Into the remaining two vials 100 µl of the 787 ng/ml 2,3,7,8-TCDD check solution is added. Each vial is evaporated to dryness using filtered nitrogen then redissolved with 100 µl of toluene. All four samples are analyzed using the above conditions and results are calculated as follows:

$$C_{ss} = C_{chk} \times \frac{(A_{ss1}/A_{ssis1}) + (A_{ss2}/A_{ssis2})}{(A_{chk1}/A_{chkis1}) + (A_{chk2}/A_{chkis2})}$$

C_{ss} = Concentration of 2,3,7,8-TCDD in stock standard solution (ng/ml).

C_{chk} = Concentration of 2,3,7,8-TCDD in check solution (ng/ml).

A_{ss1} = SIM response for 2,3,7,8-TCDD in stock standard vial 1 m/z (322 + 320).

A_{ssis1} = SIM response for $^{13}C_{12}$-2,3,7,8-TCDD in stock standard vial 1 m/z (334 + 332).

A_{ss2} = SIM response for 2,3,7,8-TCDD in stock standard vial 2 m/z (322 + 320).

A_{ssis2} = SIM response for $^{13}C_{12}$-2,3,7,8-TCDD in stock standard vial 2 m/z (334 + 332).

A_{chk1} = SIM response for 2,3,7,8-TCDD in check solution vial 1 m/z (322 + 320).

A_{chkis1} = SIM response for $^{13}C_{12}$-2,3,7,8-TCDD in check solution vial 1 m/z (334 + 332).

A_{chk2} = SIM response for 2,3,7,8-TCDD in check solution vial 2 m/z (322 + 320).

A_{chkis2} = SIM response for $^{13}C_{12}$-2,3,7,8-TCDD in check solution vial 2 m/z (334 + 332).

The five stock standard concentrations are adjusted using the verified concentration determined above.

B. Preparation of Calibration Standards

Twenty calibration standards are prepared to cover the entire range of analyses. A 5 µl aliquot of n-tetradecane is added to each of 20 silanized 1 ml cone shaped vials. Ten of the vials are spiked with 10 ng of $^{13}C_{12}$-2,3,7,8-TCDD by the addition of 100 µl of the 100 ng/ml spiking solution. The remaining ten vials are spiked with 100 ng of $^{13}C_{12}$-2,3,7,8-TCDD by the addition of 100 µl of the 1.0 µg/ml spiking solution[1]. Various quantities of native 2,3,7,8-TCDD are then added as listed in the following table.

Std.	$^{13}C_{12}$-TCDD Amount (ng)	TCDD 2 Amount (ng)	Stock Standard TCDD Soln. 2 Vol. (µl)	Stock Standard TCDD Soln. 2 Conc. (ng/ml)
10.1	10	0.5	100	5.0
10.2	10	1.0	200	5.0
10.3	10	2.0	400	5.0
10.4	10	5.0	100	50.0
10.5	10	10.0	200	50.0
10.6	10	20.0	400	50.0
10.7	10	50.0	100	500.0
10.8	10	100.0	200	500.0
10.9	10	200.0	400	500.0
10.10	10	500.0	100	5000
100.1	100	5.0	100	50.0
100.2	100	10.0	200	50.0
100.3	100	20.0	400	50.0
100.4	100	50.0	100	500.0
100.5	100	100.0	200	500.0
100.6	100	200.0	400	500.0
100.7	100	500.0	100	5000
100.8	100	1000	200	5000
100.9	100	2000	400	5000
100.10	100	5000	100	50,000

Each vial is mixed by sonication and is evaporated to dryness using filtered nitrogen. The samples must be dried to remove acetone. The series 10 samples are redissolved in 50 µl of toluene and the 100 series samples in 200 µl toluene.

1 It is essential that the same $^{13}C_{12}$-TCDD spiking solutions are used in the preparation of calibration standards and the spiking of samples.

2 The verified concentration of the stock standard 2,3,7,8-TCDD solutions must be used to calculate the amount of native TCDD actually in each vial.

C. Instrument Calibration

A minimum of four calibration standards in the expected range of analysis are analyzed each day. Standards containing 10 ng of the $^{13}C_{12}$-2,3,7,8-TCDD spike are used only for the analysis of samples spiked with 10 ng of $^{13}C_{12}$-2,3,7,8-TCDD and similarly standards containing 100 ng of the $^{13}C_{12}$-2,3,7,8-TCDD spike are used only for the analysis of samples spiked with 100 ng of $^{13}C_{12}$-2,3,7,8-TCDD. Injections of 1.6 µl are performed and peak area responses or peak heights are tabulated for ions 334, 332, 322, and 320. The ratio of SIM response m/z (322 + 320)/(334 + 332) is calculated for each standard and a calibration curve is determined by linear regression.

For X = Ratio of SIM response m/z (322 + 320)/(334 + 332)
Y = ng of Native 2,3,7,8-TCDD
n = Number of data points

$$Y = a + b(X)$$

where

$$b = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n\sum X_i^2 - (\sum X_i)^2}$$

$$a = \frac{\sum Y_i}{n} - \frac{b\sum X_i}{n}$$

Correlation Coefficient

$$r = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{[n\sum X_i^2 - (\sum X_i)^2][n\sum Y_i^2 - (\sum Y_i)^2]}}$$

Standard Error of Estimate

$$S_e = \sqrt{\frac{[Y_i - (a + b(X_i))]^2}{n-2}}$$

% Relative Difference (Y_i)

$$\%\ \text{Rel. Diff.}(Y) = \frac{2[Y_i - (a + b(X_i))]}{Y_i + a + b(X_i)} \times 100$$

The correlation coefficient is determined for each calibration. To ensure the straightness of the line the correlation coefficient must be greater than 0.990. In addition the percent relative difference between Y (regression) and Y (actual) is calculated for each point. The % rel. diff. (Y) must be within ±10% for each point to ensure precision over the working range of analysis. If these conditions are not met a recalibration must be performed by additional injections and if necessary using fresh calibration standards. A sample calibration curve is illustrated in Figure I. Figure II is an enlargement of the same curve at the low end of analysis.

DATA REPORTING

Results are reported in units of (ng/g) or parts per billion (ppb). Three significant figures are reported for values above 10.0 ppb and two significant figures fr values below 10 ppb. In addition the data package must include the following:

1. Sample log number and identification number.
2. The weight of the original wet sample aliquot.
3. The calculated value for native 2,3,7,8-TCDD.
4. If no 2,3,7,8-TCDD was detected, ND is reported with the calculated detection limit in parentheses.
5. Analytical date.
6. The response ratios of 320/322 and 332/334.
7. The results of method blanks.
8. The results of duplicate analyses.
9. The percent recovery of native TCDD from samples spiked near detection limits.
10. Daily calibration report.
11. The mass chromatograms for all samples and standards including the raw peak response data for ions 320, 322, 332, and 334.
12. The mass chromatograms for the determination of isomer specificity.
13. The mass chromatograms for qualitative verifications of identity including data for ions 194, 196, 257, 259, 320, 322 and 324.
14. Documentation on the source of the native and labeled 2,3,7,8-TCDD standards used.

Drafted by:

Date: 1-9-84

Approved by:

Date: Jan 9, 1984

LINEAR REGRESSION

X = Ratio of SIM Response m/z (322+320)/(334+332)
Y = ng Native TCDD

NO.	X	Y	Y(Reg)	% Rel Diff (Y)
1	.0629	.728	.67639	7.34977
2	.1158	1.456	1.42788	1.94982
3	.2603	3.64	3.48064	4.47589
4	.5302	7.28	7.31483	-.477275
5	1.0684	14.56	14.9605	-2.71316
6	2.6081	36.4	36.8334	-1.18351
7	5.0731	72.8	71.851	1.3121
8	10.287	145.6	145.919	-.219128

Y = 14.2059 (X) + -.217164

STANDARD ERROR OF ESTIMATE = .479726
COEFFICIENT OF DETERMINATION = .999924
CORRELATION COEFFICIENT = .999962

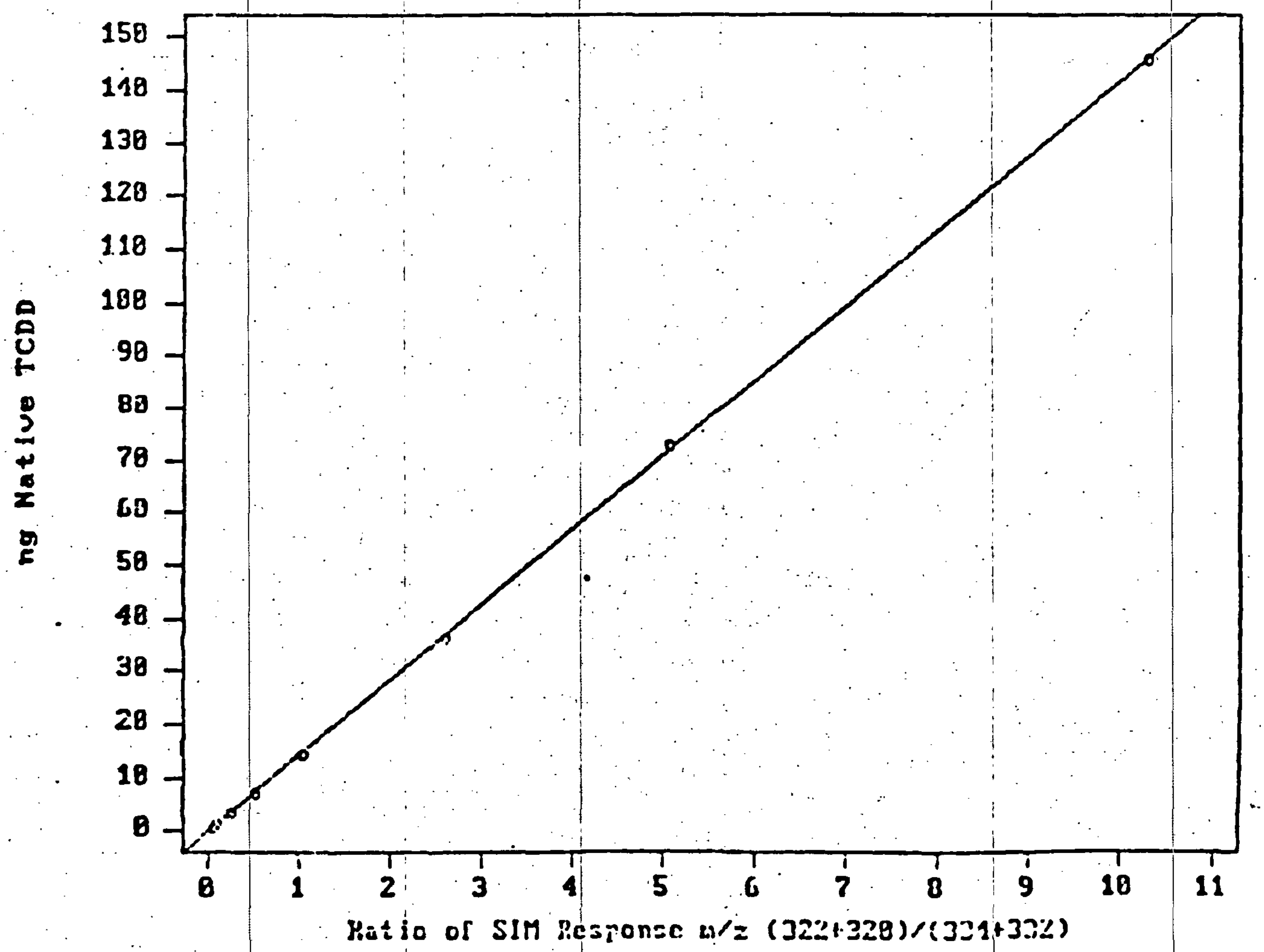


FIGURE I

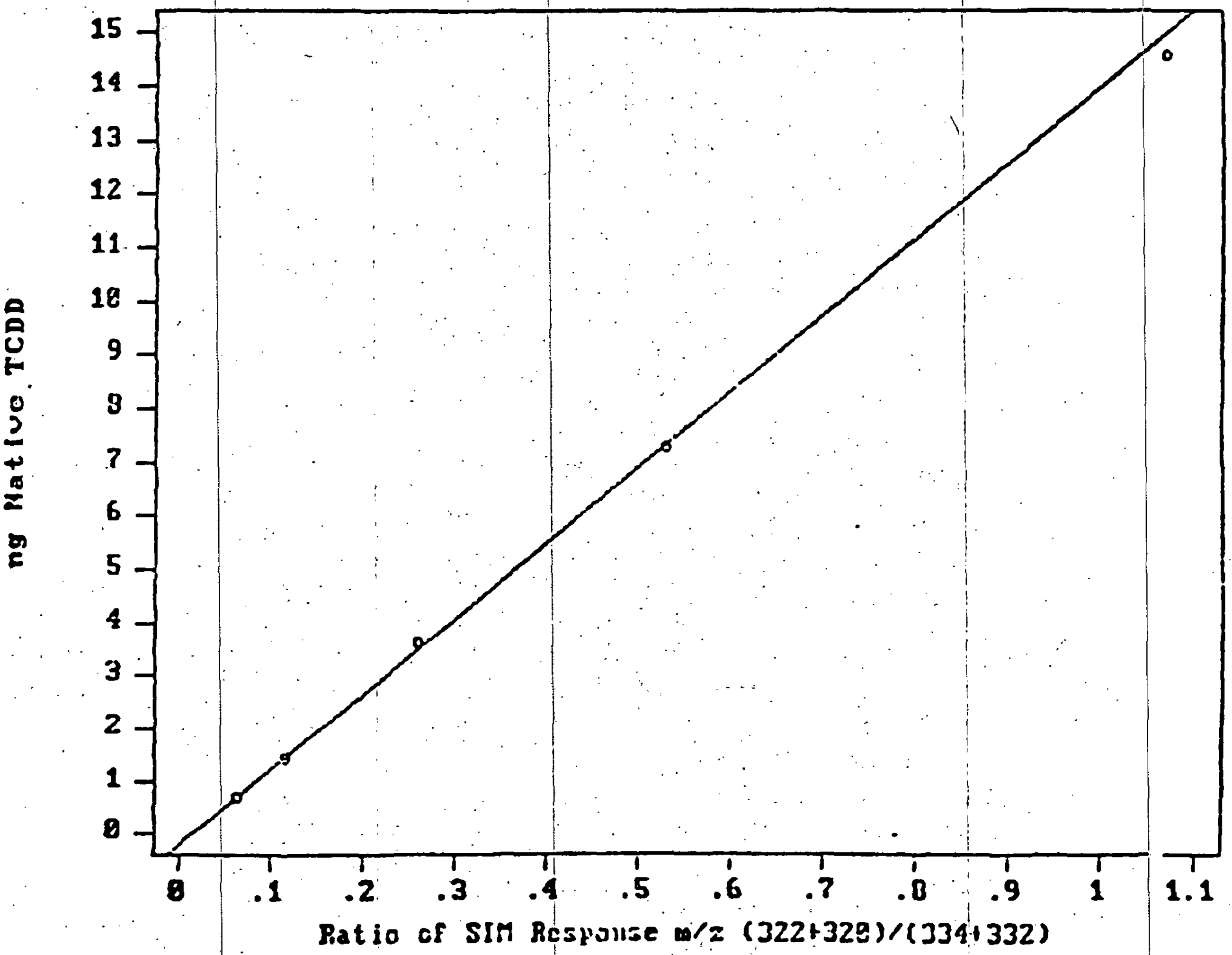

15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
ng Native TCDD
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
1.1
Ratio of SIM Response m/z (322+328)/(334+332)

FIGURE II

Sample Analysis

A. Samples are analyzed using the conditions above. The SIM area response for ions 334, 332, 322, and 320 are tabulated and the ratio of SIM response m/z (322 + 320)/(334 + 332) is calculated. Results are determined as follows:

Where X_s = Ratio of SIN response m/z (322 + 320)/(334 + 332) for the sample.
W = Weight of sample (wet) (g)

$$\text{Result (ppb)} = \frac{a + b\,(X_s)}{W}$$

a and b are determined by regression analysis of the calibration standards.

B. In cases where no native 2,3,7,8-TCDD is detected, the actual detection limit is calculated based on a signal to noise ratio of 2.5 to 1 for both ions 320 and 322. The detection limit is calculated by comparison of peak heights using calibration standard 10.2 for samples spiked with 10 ng $^{13}C_{12}$-2,3,7,8-TCDD and calibration standard 100.2 for samples spiked with 100 ng of $^{13}C_{12}$-2,3,7,8-TCDD. The noise level is measured for ions 320 and 322 in the sample. The peak heights for ions 332 and 334 are also measured in the sample. The peak heights for 320, 322, 332, and 334 are then measured in the calibration standard. The limit is then calculated as follows:

$$\text{Det. Limit} = \frac{A_{CS} \times 2.5 \times H_S \times H_{RCS}}{W \times H_{CS} \times H_{RS}}$$

Where

A_{CS} = Amount of native TCDD in calibration std. (ng)
W = Weight of sample (wet) (g)
H_S = (Height of sample noise 320 + 322) X plotter scale
H_{CS} = (Height of calib. std. 320 + 322) X plotter scale
H_{RS} = (Height of sample 332 + 334) X plotter scale
H_{RCS} = (Height of calib. std. 332 + 334) X plotter scale

Note: Ions 320 and 322 are plotted on same scale.
Ions 332 and 334 are plotted on same scale.

If an interfering signal is present at 320 or 322 the ion not intefered with is used to calculate a detection limit (using responses 320 and 332 <u>or</u> 322 and 334). If both ions are subject to interferences which are more than 2.5 times the noise level, the detection limit is calculated using the summed interference levels of both 320 and 322 without multiplication by 2.5.

QUALITY CONTROL

1. A laboratory 'method blank' must be run with each set of 24 or fewer samples. The method blank is prepared as a sample without the introduction of soil, sediment or sludge. It is spiked prior to extraction with the 10 ng spiking solution.

2. Performance evaluation samples, available through the EPA, must be analyzed periodically. If the performance criteria are not met appropriate corrective action must be taken and demonstrated before sample analysis is continued.

3. A minimum of one per set of 20 or less samples is analyzed in duplicate.

4. A minimum of one blank sample per set must be spiked with native 2,3,7,8-TCDD at a level near the detection limit required for that set of samples.

5. For each sample the internal standard must be present with a minimum signal to noise ratio of 10 to 1 for both ions 334 and 332. In addition the ion ratio of 332/334 must be within 0.67 to 0.87. If these conditions are not met further sample clean-up and possibly reextraction of a fresh sample are required.

6. Qualitative identification of 2,3,7,8-TCDD is performed using the following guidelines.

   A. Isomer specificity is determined at least once with each set of 20 or less samples. The determination consists of injecting a mixture of $^{13}C_{12}$-2,3,7,8-TCDD and seven TCDD isomers (2378, 1478, 1234, 1237, 1238, 1278, 1267) using the analytical conditions. The degree of isomer specificity is determined by calculating the % valley relative to the 2,3,7,8-TCDD peak height. An example of this determination is illustrated in Figure III.

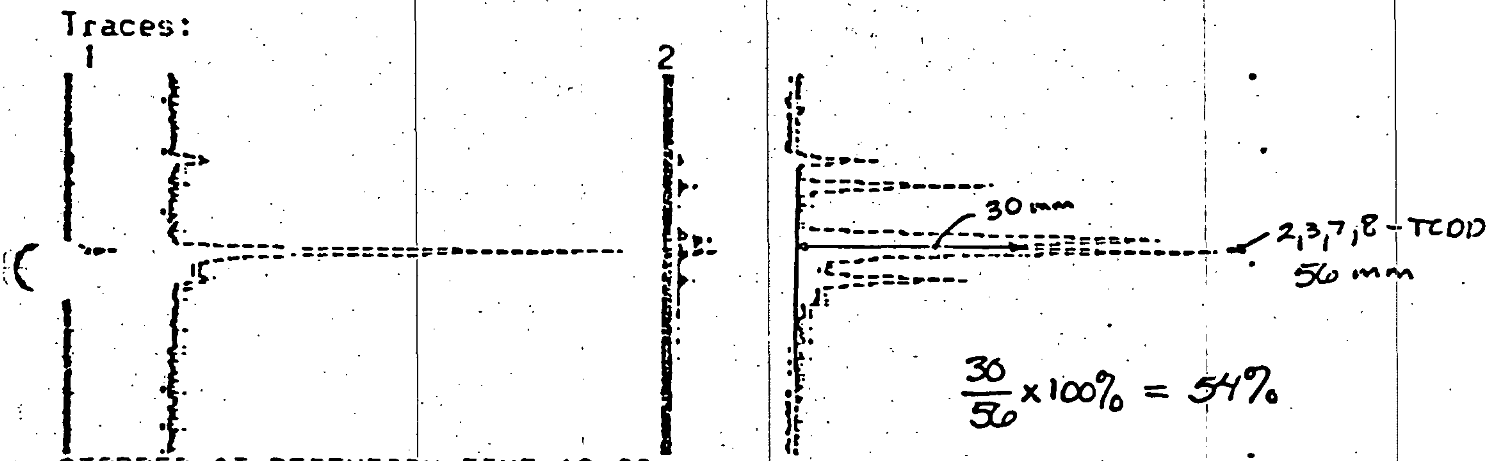
   B. The ratio of ions 320/322 must be within 0.67 to 0.87. If this condition is not met then the sample must be subjected to additional clean-up.

   C. Ions 320, 322, 332, and 334 must all maximize togethey within 0.03 min. of one another, and ions 320 and 322 must both be present at a level greater than 2.5 x the noise level.

   D. At least one positive sample per set of 20 or less must be analyzed monitoring ions 194, 196, 257, 259, 320, 322, 324 where the following expected ion ratio are verified:

      320/322 = 0.67 to 0.87
      320/324 = 1.42 to 1.74
      257/259 = .93 to 1.13
      194/196 = 1.39 to 1.69

```
********************** SIM ION RECONSTRUCT (Rev. 1/21/83) **********************

ION MASSES              334.00   332.00   321.90   319.90
DWELL TIMES (msec)       60.00    50.00    60.00    60.00
MAX PEAK HEIGHTS          6.40     5.51     5.04     4.48

FILE SIM  7    TOTAL RUN TIME =15.91   SAMPLE SIZE =     1.60

DATA IS NOT SMOOTHED

              Plot from Ret Time 11.00  of Disc 07
                   to Ret Time  12.99

           Data Acquired on   1/05/1983 12:57 PM
           |              Plot Summary                |
           | Trace |     Ion Mass     |   Full Scale  |
           |   1   |   334.00 amu     |      6.85     |
           |   2   |   321.90 amu     |      5.41     |
```



$\frac{30}{56} \times 100\% = 54\%$

```
STOPPED AT RETENTION TIME 12.99
```

FIGURE III

7 Isomer Mixture Plus $^{13}C_{12}$-2,3,7,8-TCDD

Attachment D

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

DATE December 23, 1983

SUBJECT. Fish and Sediment Sampling on the Spring River

FROM Daniel J. Harris
Environmental Engineer, EP&R/ENSV

TO Robert L. Morby
Chief, WMBR/ARWM

THRU: William J. Keffer
Chief, EP&R/ENSV

John C. Wicklund
Director, ENSV

David A. Wagoner
Director, ARWM

In response to your request, two of the six stations specified in the Syntex draft protocol for the Spring River were sampled on December 15, 1983.

Those individuals in attendance for this effort included the following:

Ron Crunkleton, Missouri Department of Conservation
James Civielli, Missouri Department of Conservation
Howard Kerns, Missouri Department of Conservation
Glen Davis, Syntex Representative
Bob Wiggins, FIT
Dan Harris, EPA/EP&R

Information on the two stations and the samples collected is as follows:

LOCATION: 12 miles downstream from Syntex

SAMPLES: Fish and Sediment

(Fish)

TIME OF COLLECTION: 1215 to 1330 hours

COLLECTED BY: Crunkleton, Civielli, Kerns

DESCRIPTION OF FISH: Species, Hog Suckers

NUMBER OF FISH, LENGTH AND WEIGHT:

No.	Length, Inches	Weight, Kilograms
1	11.5	0.28
2	11.3	0.32
3	12.2	0.42
4	12.6	0.42
5	11.6	0.30
6	15.5	0.86
7	14.9	0.58
8	13.2	0.42
9	13.2	0.42
10	11.2	0.30

ASSIGNED COMPOSITE LABORATORY NUMBER: AAC401

(Sediment)

TIME OF COLLECTION: 1245 hours

COLLECTED BY: Wiggins, Davis

DEPTH OF COLLECTION: 0 to 6 inches

METHOD: Using spoons and pole with attached cup sediment was transferred to a stainelss-steel pan where it was blended and mixed prior to transfer to separate laboratory containers.

ASSIGNED SAMPLE NUMBER: AAC400

SIZE OF SAMPLE CONTAINER: 1-quart glass jar

LOCATION: 0.3 miles downstream from Syntex

SAMPLES: Fish and Sediment

(Fish)

TIME OF COLLECTION: 1600 to 1615 hours

COLLECTED BY: Crunkleton, Civielli, Kerns

DESCRIPTION OF FISH: Species - White Suckers

NUMBER OF FISH, LENGTH AND WEIGHT:

No.	Length, Inches	Weight, Kilograms
1	11.0	0.24
2	12.8	0.40
3	11.7	0.30
4	11.6	0.30
5	12.9	0.32
6	13.2	0.40
7	13.9	0.48
8	15.2	0.60
9	13.9	0.48
10	15.5	0.68

ASSIGNED COMPOSITE LABORATORY NUMBER: AAC403

(Sediment)

TIME OF COLLECTION: 1615 hours

COLLECTED BY: Harris, Wiggins

DEPTH OF COLLECTION: 0 to 6 inches

METHOD: Using pole with attached cup sediment was transferred to a stainless-steel pan where it was blended and mixed prior to transfer to separate laboratory containers.

ASSIGNED SAMPLE NUMBER: AAC402

SIZE OF SAMPLE CONTAINER: 1-quart glass jar

Splits of the two sediment samples were turned over to Glen Davis (Syntex) in the field at about 1700 hours December 15, 1983.

The two sets of fish samples were returned in their entirety to the Regional Laboratory.

As per conversation with Ron Crunkleton, fish sample AAC401 (12 mile station) will be analyzed for TCDD by compositing the fillets of each of the ten fish with skin off.

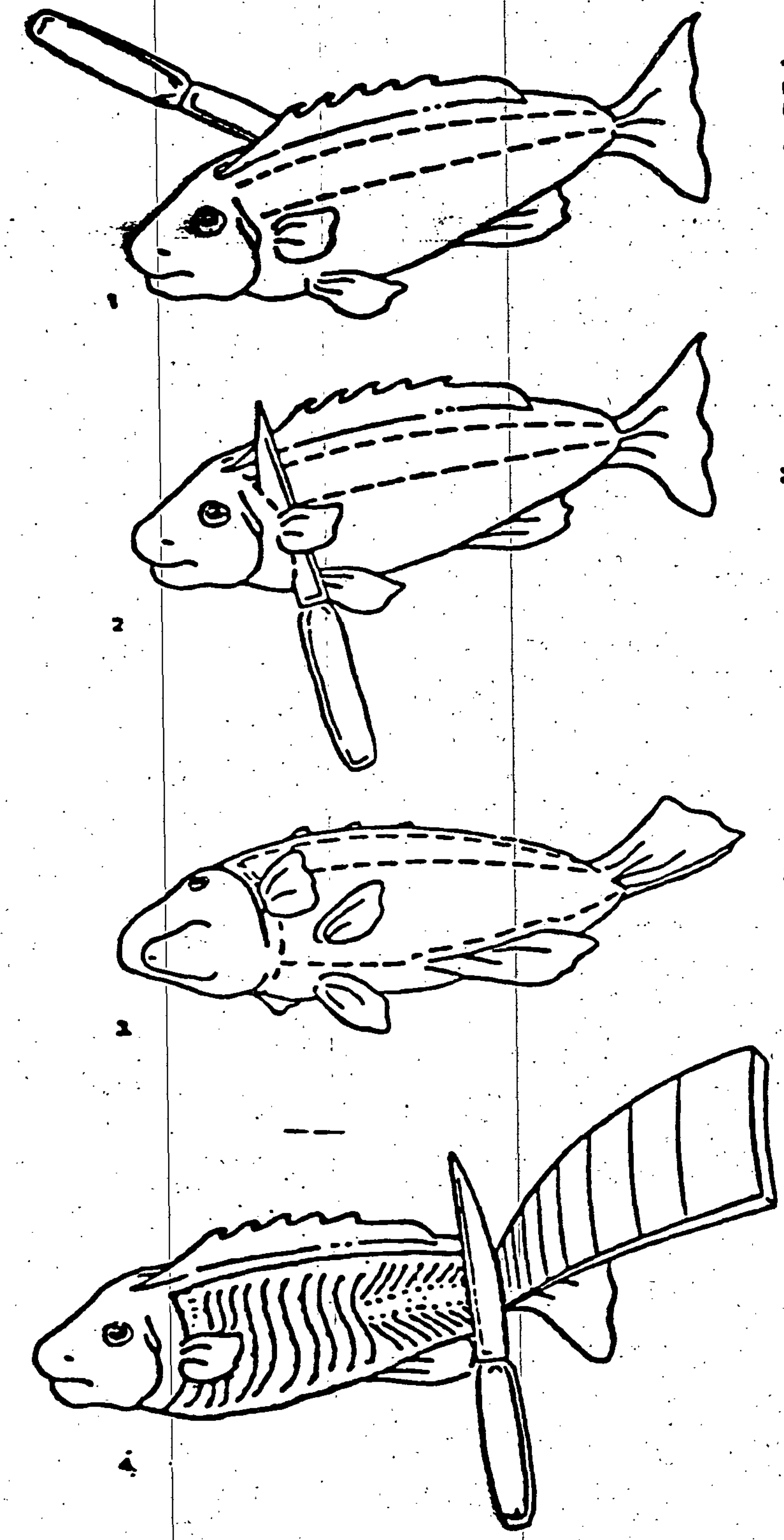
Fish sample AAC403 (0.3 mile station) will be analyzed in two ways. The fillets of all the fish with skin off will be combined to make up one sample. For the other sample, the remainder of the fish including the skin will be combined. Splist of the three homogenates will be provided to Syntex.

Ron Crunkelton took scale samples of the fish for examination in the laboratory to determine the respective ages of the fish. He should be providing you with this information shortly.

I shall leave it up to you to provide Syntex a copy of this memorandum on our field effort.

cc: Scott Ritchey, ARWM
Charlie Hensley, LABO/ENSV
Bob Kleopfer, LABO/ENSV

Attacnment E



THE FOLLOWING PROCEDURE IS USED TO PRODUCE THE "STANDARD FILLET":

1. MAKE A SHALLOW CUT THROUGH THE SKIN (ON EITHER SIDE OF THE DORSAL FIN) FROM BASE OF THE HEAD TO THE TAIL.

2. MAKE A CUT BEHIND THE ENTIRE LENGTH OF THE GILL COVER CUTTING THROUGH SKIN AND FLESH TO THE BONE.

3. MAKE A CUT ALONG THE BELLY FROM THE BASE OF THE PECTORAL FIN TO THE TAIL AS SHOWN.

4. REMOVE THE FILLET AND REMOVE THE MAJOR BONES.

Figure 3-12 FISH FILLET PROCEDURE

Attachment [illegible]

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

DATE: December 23, 1983

SUBJECT: Sediment Sampling of the Spring River under the Syntex Consent Agreement

FROM: Daniel J. Harris DJH/wm
Environmental Engineer, EP&R/ENSV

TO: Scott Ritchey, ARWM

THRU: William J. Keffer wjk/wm
Chief, EP&R/ENSV

John C. Wicklund
Director, ENSV

David A. Wagoner
Director, ARWM

From the December 8, 1983, meeting with Syntex, it is my understanding that Syntex is more or less content to have us specify the methods for sediment collection and handling, since we will be physically doing the collection in conjunction with Crunkleton's collection of the fish.

Based upon that understanding, I am providing the following specific details to clarify those points raised by you at the meeting:

1. EPA and Syntex samples will be placed in 16-ounce glass jars with Teflon lid liners. The jars shall be overpacked in 1-gallon paint cans with appropriate tags and labels. All packaging will be provided by ENSV.

2. Under method of collection, I would only specify that sediment will be collected manually from a nominal 0 to 6-inch depth and that collection of samples will, to the extent possible, be similar at each of the six stations throughout the period of monitoring. Sediment will be collected across the channel.

3. ENSV will prepare a report following each yearly collection effort. This report will include laboratory sample numbers, date and time of collection, method of collection and parties in attendance. In addition, Syntex will be notified in advance of the date or dates of collection and will be invited to send a representative to observe, take notes, photographs, etc.

4. The sample splits for Syntex will be hand delivered to any Syntex representative specified by Ray Forrester and ENSV shall request a receipt.

5. In preparing the sample spilts, the aliquots from the stream bed will be initially transferred to clean stainless-steel pans where the material will be thoroughly mixed and blended prior to dividing between the two laboratory sample containers.

ATTACHMENT G

Applied Regression Analysis

N. R. DRAPER
University of Wisconsin

H. SMITH
The Procter & Gamble Company

John Wiley & Sons, Inc.
New York · London · Sydney

From the regression line of weight on height we could find an average observed weight of individuals of the given height and use this as an estimate of the weight that we did not record.

[A pair of random variables such as (height, weight) follows some sort of bivariate probability distribution. When we are concerned with the dependence of a random variable Y on a quantity X which is variable but *not* a random variable, an equation that relates Y to X is usually called a *regression equation*. Although the name is, strictly speaking, incorrect, it is well established and conventional.]

We can see that whether a relationship is exactly linear or linear only insofar as mean values are concerned, knowledge of the relationship will be useful. (The relationship might, of course, be more complicated than linear but we shall consider this later.)

A linear relationship may be a valuable one even when we *know* that a linear relationship cannot be true. Consider the response relationship shown in Figure 1.2. It is obviously not linear over the range $0 \leq X \leq 100$. However, if we were interested primarily in the range $0 \leq X \leq 45$, a straight-line relationship evaluated from observations in this range might provide a perfectly adequate representation of the function *in this range*. The relationship thus fitted would, of course, not apply to values of X outside this restricted range and could not be used for predictive purposes outside this range.

(Similar remarks can be made when more than one independent variable is involved. Suppose we wish to examine the way in which a response Y depends on variables $X_1, X_2, \ldots, X_k$. We determine a regression equation from data which "cover" certain areas of the "X-space." Suppose the point $\mathbf{X}_0 = (X_{10}, X_{20}, \ldots, X_{k0})$ lies *outside* the regions covered by the original data. While we can mathematically obtain a predicted value $\hat{Y}(\mathbf{X}_0)$ for the response at the point $\mathbf{X}_0$, we must realize that reliance on such a prediction is extremely dangerous and becomes more dangerous the further $\mathbf{X}_0$ lies from the original regions, unless some additional knowledge is available that the regression equation is valid in a wider region of the X-space. Note that it is sometimes difficult to realize at first that a suggested point lies outside a region in a multi-dimensional space. To take a simple example consider the region defined by the ellipse in Figure 1.3. We see that there are points in the region for which $1 \leq X_1 \leq 9$ and for which $2.4 \leq X_2 \leq 6.3$. Although both coordinates of P lie within these ranges, P itself lies outside the region. When more dimensions are involved misunderstandings of this sort easily arise.)

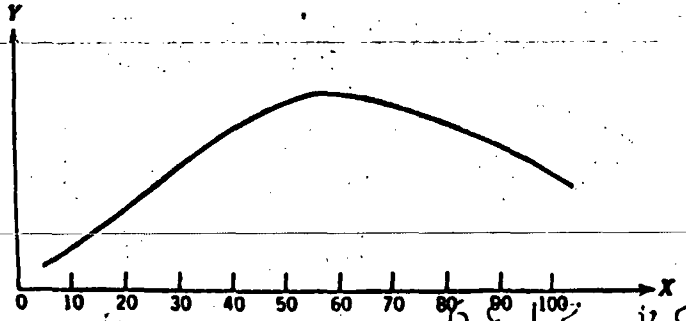


Figure 1.2

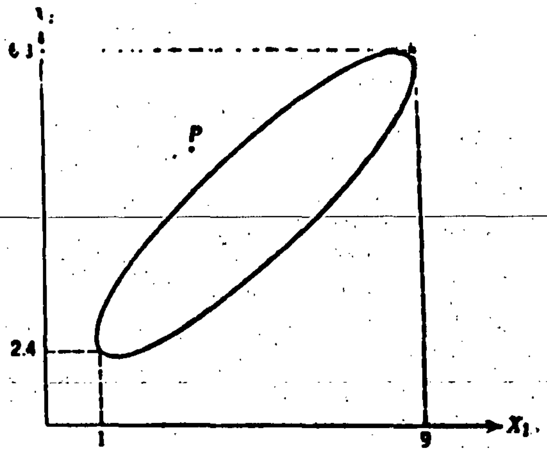


Figure 1.3

1.2. Linear Regression: Fitting a Straight Line

We have mentioned that in many situations a straight-line relationship can be valuable in summarizing the observed dependence of one variable on another. We now show how the equation of such a straight line can be obtained by the method of least squares when data are available. Consider, in the printout on page 352, the 25 observations of variable 1 (pounds of steam used per month) and variable 8 (average atmospheric temperature in degrees Fahrenheit). The corresponding pairs of observations are given in Table 1.1 and are plotted in Figure 1.4.

Let us tentatively assume that the regression line of variable 1 which we shall denote by Y, on variable 8(X) has the form $\beta_0 + \beta_1 X$. Then we can write the linear, first-order model

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{1.2.1}$$

Table 1.1 Twenty-five Observations of Variables 1 and 8

Observation Number	Variable Number 1(Y)	8(X)
1	10.98	35.3
2	11.13	29.7
3	12.51	30.8
4	8.40	58.8
5	9.27	61.4
6	8.73	71.3
7	6.36	74.4
8	8.50	76.7
9	7.82	70.7
10	9.14	57.5
11	8.24	46.4
12	12.19	28.9
13	11.88	28.1
14	9.57	39.1
15	10.94	46.8
16	9.58	48.5
17	10.09	59.3
18	8.11	70.0
19	6.83	70.0
20	8.88	74.5
21	7.68	72.1
22	8.47	58.1
23	8.86	44.6
24	10.36	33.4
25	11.08	28.6

that is, for a given X, a corresponding observation Y consists of the value $\beta_0 + \beta_1 X$ plus an amount ϵ, the increment by which any individual Y may fall off the regression line. Equation (1.2.1) is the *model* of what we believe. We begin by assuming that it holds; but we shall have to inquire at a later stage if indeed it does. In many aspects of statistics it is necessary to assume a mathematical model to make progress. It might be well to emphasize that what we are usually doing is to *consider* or *tentatively entertain* our model. The model must always be critically examined somewhere along the line. It is our "opinion" of the situation at one stage of the investigation and our "opinion" must be changed if we find, at a later stage, that the facts are against it. β_0 and β_1 are called the *parameters* of the model.

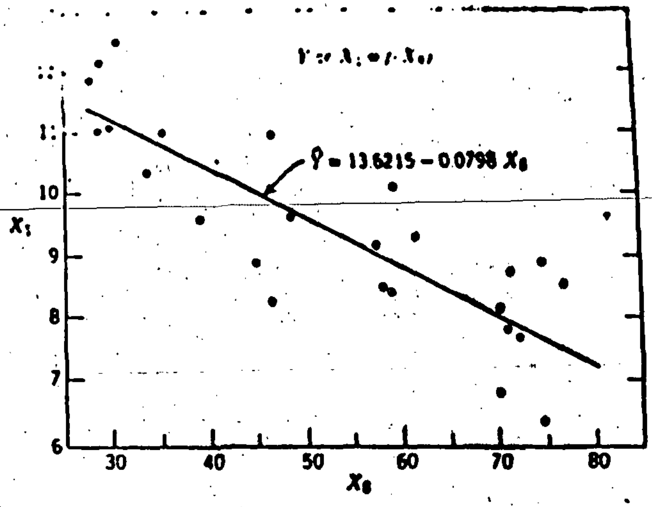


Figure 1.4

(*Note:* When we say that a model is linear or nonlinear, we are referring to linearity or nonlinearity *in the parameters*. The value of the highest power of an independent variable in the model is called the *order* of the model. For example,

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon$$

is a second-order (in X) linear (in the β's) regression model. Unless a model is specifically called nonlinear it can be taken that it is linear in the parameters, and the word linear is usually omitted and understood. The order of the model could be of any size.)

Now β_0, β_1, and ϵ are unknown in Eq. (1.2.1), and in fact ϵ would be difficult to discover since it changes for each observation Y. However, β_0 and β_1 remain fixed and, although we cannot find them exactly without examining all possible occurrences of Y and X, we can use the information provided by the twenty-five observations in Table 1.1 to give us *estimates* b_0 and b_1 of β_0 and β_1; thus we can write

$$\hat{Y} = b_0 + b_1 X, \tag{1.2.2}$$

where $\hat{Y}$, read "Y hat," denotes the *predicted* value of Y for a given X, when b_0 and b_1 are determined. Equation (1.2.2) could then be used as a predictive equation; substitution for a value of X would provide a prediction of the true mean value of Y for that X.

Our estimation procedure will be that of *least squares*. Under certain assumptions to be mentioned later, this procedure has certain properties.

For the moment we state it as our chosen method of estimating the parameters without justification. Suppose we have available n sets of observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. (In our example $n = 25$.) Then by Eq. (1.2.1) we can write

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad (1.2.3)$$

so that the sum of squares of deviations from the true line is

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2. \qquad (1.2.4)$$

We shall choose our estimates b_0 and b_1 to be the values which, when substituted for β_0 and β_1 in Eq. (1.2.4), produce the least possible value of S. (Note that X_i, Y_i are the fixed numbers which we have observed.) We can determine b_0 and b_1 by differentiating Eq. (1.2.4) first with respect to β_0 and then with respect to β_1 and setting the results equal to zero. Now

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) \end{aligned} \qquad (1.2.5)$$

so that the estimates b_0 and b_1 are given by

$$\begin{aligned} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) &= 0 \\ \sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) &= 0 \end{aligned} \qquad (1.2.6)$$

where we substitute (b_0, b_1) for (β_0, β_1), when we equate Eq. (1.2.5) to zero. From (1.2.6) we have

$$\begin{aligned} \sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i &= 0 \\ \sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 &= 0 \end{aligned} \qquad (1.2.7)$$

or

$$\begin{aligned} b_0 n + b_1 \sum_{i=1}^{n} X_i &= \sum_{i=1}^{n} Y_i \\ b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i \end{aligned} \qquad (1.2.8)$$

These equations are called the *normal equations*.

The solution of Eq. (1.2.8) for b_1 is

$$b_1 = \frac{\sum X_i Y_i - [(\sum X_i)(\sum Y_i)]/n}{\sum X_i^2 - (\sum X_i)^2/n} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \qquad (1.2.9)$$

where all summations are from $i = 1$ to n and the two expressions for b_1 are just slightly different forms of the same quantity since, defining

$$\begin{aligned} \bar{X} &= (X_1 + X_2 + \cdots + X_n)/n = \sum X_i/n, \\ \bar{Y} &= (Y_1 + Y_2 + \cdots + Y_n)/n = \sum Y_i/n, \end{aligned}$$

we have that

$$\begin{aligned} \sum (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum X_i Y_i - \bar{X} \sum Y_i - \bar{Y} \sum X_i + n\bar{X}\bar{Y} \\ &= \sum X_i Y_i - n\bar{X}\bar{Y} \\ &= \sum X_i Y_i - (\sum X_i)(\sum Y_i)/n. \end{aligned}$$

The first form in Eq. (1.2.9) is normally used when actually computing the value of b_1. The solution of Eq. (1.2.8) for b_0 is

$$b_0 = \bar{Y} - b_1 \bar{X} \qquad (1.2.10)$$

The quantity $\sum X_i^2$ is called the *uncorrected sum of squares of the X's*, and $(\sum X_i)^2/n$ is the *correction for the mean of the X's*. The difference is called the *corrected sum of squares of the X's*. Similarly, $\sum X_i Y_i$ is called the *uncorrected sum of products*, and $(\sum X_i)(\sum Y_i)/n$ is the *correction for the means*. The difference is called the *corrected sum of products of X and Y*. Substituting Eq. (1.2.10) into Eq. (1.2.2) gives the estimated regression equation

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X}), \qquad (1.2.11)$$

where b_1 is given by Eq. (1.2.9). Let us now perform these calculations on the data given as an example in Table 1.1. We find the following:

$$\begin{aligned} n &= 25 \\ \sum Y_i &= 10.98 + 11.13 + \cdots + 11.08 = 235.60 \\ \bar{Y} &= 235.60/25 = 9.424 \\ \sum X_i &= 35.3 + 29.7 + \cdots + 28.6 = 1315 \\ \bar{X} &= 1315/25 = 52.60 \\ \sum X_i Y_i &= (10.98)(35.3) + (11.13)(29.7) + \cdots + (11.08)(28.6) \\ &= 11821.4320 \\ \sum X_i^2 &= (35.3)^2 + (29.7)^2 + \cdots + (28.6)^2 = 76323.42 \end{aligned}$$

$$b_1 = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n}$$

$$b_1 = \frac{11821.4320 - (1315)(235.60)/25}{76323.42 - (1315)^2/25} = \frac{-571.1280}{7154.42}$$

$$b_1 = -0.079829.$$

Table 1.2 Fitted Values, Observations, and Residuals

Observation Number	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
1	10.98	10.81	0.17
2	11.13	11.25	−0.12
3	12.51	11.17	1.34
4	8.40	8.93	−0.53
5	9.27	8.72	0.55
6	8.73	7.93	0.80
7	6.36	6.98	−1.32
8	8.50	7.50	1.00
9	7.82	7.98	−0.16
10	9.14	9.03	0.11
11	8.24	9 92	−1.68
12	12.19	11.32	0.87
13	11.88	11.38	0.50
14	9.57	10.50	−0.93
15	10.94	9.89	1.05
16	9.58	9.75	−0.17
17	10.09	8.89	1.20
18	8.11	8.04	0.07
19	6.83	8.04	−1.21
20	8.88	7.68	1.20
21	7.68	7.87	−0.19
22	8.47	8.98	−0.51
23	8.86	10.06	−1.20
24	10.36	10.96	−0.60
25	11.08	11.34	−0.26

The fitted equation is thus

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X})$$
$$\hat{Y} = 9.4240 - 0.079829(X - 52.60)$$
$$\hat{Y} = 13.623005 - 0.079829X.$$

The fitted regression line is plotted in Figure 1.4. We can tabulate for each of the 25 values X_i, at which a Y_i observation is available, the fitted value $\hat{Y}_i$ and the *residual* $Y_i - \hat{Y}_i$ as in Table 1.2. The residuals are given to the same number of places as the original data.

Note that since $\hat{Y}_i = \bar{Y} + b_1(X_i - \bar{X})$,

$$Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - b_1(X_i - \bar{X}),$$
$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \sum_{i=1}^{n}(Y_i - \bar{Y}) - b_1\sum_{i=1}^{n}(X_i - \bar{X}) = 0.$$

Thus the sum of the residuals should be zero. In fact, it is −0.02 here—a rounding error. The sum of residuals in any regression problem is always zero when there is a β_0 term in the model as a consequence of the first normal equation. The omission of β_0 from a model implies that the response is zero when all the independent variables are zero. This is a very strong assumption which is usually unjustified. In a straight-line model $Y = \beta_0 + \beta_1 X + \epsilon$ omission of β_0 implies that the line passes through $X = 0$, $Y = 0$—that is, that the line has a zero *intercept* $\beta_0 = 0$ at $X = 0$. We note here, before the more general discussion in Section 5.4, that physical removal of β_0 from the model is always possible by "centering" the data, but that this is quite different from setting $\beta_0 = 0$. For example, if we write Eq. (1.2.1) in the form

$$Y - \bar{Y} = (\beta_0 + \beta_1\bar{X} - \bar{Y}) + \beta_1(X - \bar{X}) + \epsilon$$

or

$$y = \beta_0' + \beta_1 x + \epsilon$$

say, where $y = Y - \bar{Y}$, $\beta_0' = \beta_0 + \beta_1\bar{X} - \bar{Y}$, $x = X - \bar{X}$, then the least-squares estimates of β_0' and β_1 are given as follows:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

identical to Eq. (1.2.9), while

$$b_0' = \bar{y} - b_1\bar{x} = 0, \qquad \text{since } \bar{x} = \bar{y} = 0,$$

whatever the value of b_1. Because this always happens, we can write the centered model as

$$Y - \bar{Y} = \beta_1(X - \bar{X}) + \epsilon$$

omitting the β_0' (intercept) term entirely. We have lost one parameter but there is a corresponding loss in the data since the quantities $Y_i - \bar{Y}$, $i = 1, 2, \ldots, n$ represent only $(n - 1)$ separate pieces of information due to the fact that their sum is zero, whereas $Y_1, Y_2, \ldots, Y_n$ represent n separate pieces of information. Effectively the "lost" piece of information has been used to enable the proper adjustments to be made to the model so that the intercept term can be removed.

1.3. The Precision of the Estimated Regression

We now tackle the question of what measure of precision can be attached to our estimate of the regression line. Consider the following identity:

$$Y_i - \hat{Y}_i = Y_i - \bar{Y} - (\hat{Y}_i - \bar{Y}). \tag{1.3.1}$$

If we square both sides and sum from $i = 1$ to n, we obtain

$$\begin{aligned}\sum (Y_i - \hat{Y}_i)^2 &= \sum \{(Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})\}^2\\ &= \sum \{(Y_i - \bar{Y})^2 + (\hat{Y}_i - \bar{Y})^2 - 2(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})\}\\ &= \sum (Y_i - \bar{Y})^2 + \sum (\hat{Y}_i - \bar{Y})^2 - 2\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}).\end{aligned}$$

The third term can be rewritten as

$$\begin{aligned}&-2\sum (Y_i - \bar{Y})b_1(X_i - \bar{X}) &&\text{by (1.2.11)}\\ &= -2b_1 \sum (Y_i - \bar{Y})(X_i - \bar{X})\\ &= -2b_1^2 \sum (X_i - \bar{X})^2 &&\text{by (1.2.9)}\\ &= -2\sum (\hat{Y}_i - \bar{Y})^2 &&\text{by (1.2.11).}\end{aligned}$$

Thus

$$\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \bar{Y})^2 - \sum (\hat{Y}_i - \bar{Y})^2. \qquad (1.3.2)$$

Equation (1.3.2) can be rewritten

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2. \qquad (1.3.3)$$

Now $Y_i - \bar{Y}$ is the deviation of the ith observation from the overall mean and so the left-hand side of Eq. (1.3.3) is the sum of squares of deviations of the observations from the mean; this is shortened to *SS about the mean*, and is also the *corrected sum of squares of the Y's*. Since $Y_i - \hat{Y}_i$ is the deviation of the ith observation from its predicted or fitted value (the ith *residual*—see Chapter 3), and $\hat{Y}_i - \bar{Y}$ is the deviation of the predicted value of the ith observation from the mean, we can express Eq. (1.3.3) in words as follows:

$$\begin{matrix}\text{Sum of squares}\\ \text{about the mean}\end{matrix} = \begin{matrix}\text{Sum of squares}\\ \text{about regression}\end{matrix} + \begin{matrix}\text{Sum of squares}\\ \text{due to regression}\end{matrix}.$$

This shows that, of the variation in the Y's about their mean, some of the variation can be ascribed to the regression line and some, $\sum (Y_i - \hat{Y}_i)^2$, to the fact that the actual observations do not all lie on the regression line—if they all did, the sum of squares about the regression would be zero! From this procedure we can see that a way of assessing how useful the regression line will be as a predictor is to see how much of the SS about the mean has fallen into the SS due to regression and how much into the SS about regression. We shall be pleased if the SS due to regression is much greater than the SS about regression, or what amounts to the same thing if the ratio $R^2 = $ (*SS due to regression*)/(*SS about mean*) is not too far from unity.

Any sum of squares has associated with it a number called its degrees of freedom. This number indicates how many independent pieces of information involving the n independent numbers $Y_1, Y_2, \ldots, Y_n$ are needed to compute the sum of squares. For example, the SS about the mean needs $(n - 1)$ independent pieces (for of the numbers $Y_1 - \bar{Y}, Y_2 - \bar{Y}, \ldots, Y_n - \bar{Y}$, only $(n - 1)$ are independent since all n numbers sum to zero by definition of the mean). We can compute the SS due to regression from a single function of $Y_1, Y_2, \ldots, Y_n$, namely b_1 [since $\sum (\hat{Y}_i - \bar{Y})^2 = b_1^2 \sum (X_i - \bar{X})^2$], and so this sum of squares has one degree of freedom. By subtraction, the SS about regression has $(n - 2)$ degrees of freedom. Thus, corresponding to Eq. (1.3.3), we can show the split of degrees of freedom as

$$(n - 1) = (n - 2) + 1. \qquad (1.3.4)$$

Using Eqs. (1.3.3) and (1.3.4) and employing alternative computational forms for the expressions of Eq. (1.3.3) we can construct an *analysis of variance* table in the following form:

Source	Sum of Squares	Degrees of Freedom	Mean Square
Regression	$b_1\left\{\sum X_iY_i - \dfrac{(\sum X_i)(\sum Y_i)}{n}\right\}$	1	MS_R
About regression (residual)	By subtraction	$n - 2$	$s^2 = \dfrac{(SS)}{(n-2)}$
About mean (total, corrected for mean)	$\sum Y_i^2 - \dfrac{(\sum Y_i)^2}{n}$	$n - 1$	

The "Mean Square" column is obtained by dividing each sum of squares entry by its corresponding degrees of freedom.

A more general form of the analysis of variance table, which we do not need here but which is useful for comparison purposes later (see Section 2.2), is obtained by incorporating the correction factor for the mean of the Y's into the table where, for reasons explained in Section 2.2, it is called SS(b_0). The table takes the form:

Source	Sum of Squares	Degrees of Freedom	Mean Square
Regression (b_0)	$SS(b_0) = \dfrac{(\sum Y_i)^2}{n}$	1	
Regression ($b_1 \mid b_0$)	$SS(b_1 \mid b_0) = b_1\left\{\sum X_iY_i - \dfrac{(\sum X_i)(\sum Y_i)}{n}\right\}$	1	MS_R
Residual	By subtraction	$n - 2$	$s^2 = \dfrac{(SS)}{(n-2)}$
Total, uncorrected for mean	$\sum Y_i^2$	n	

The notation $SS(b_1 \mid b_0)$ is read "the sum of squares for b_1 after allowance has been made for b_0." The purpose of this notation is explained in Sections 2.2 and 2.7.

The mean square about regression, s^2 will provide an estimate *based on $n - 2$ degrees of freedom* of the *variance about the regression*, a quantity we shall call $\sigma^2_{Y \cdot X}$. If the regression equation were estimated from an indefinitely large number of observations, the variance about the regression would represent a measure of the error with which any observed value of Y could be predicted from a given value of X using the determined equation (see note 1 of Section 1.4).

We shall now carry out the calculations of this section for our example and then discuss a number of ways the regression equation can be examined. The SS due to regression is $b_1\{\sum X_iY_i - (\sum X_i)(\sum Y_i)/n\}$

$$= \frac{\{\sum X_iY_i - (\sum X_i)(\sum Y_i)/n\}^2}{\{\sum X_i^2 - (\sum X_i)^2/n\}}$$

$$= (-571.1280)^2/7154.42$$

$$= 45.59.$$

The Total (corrected) SS is $\sum Y_i^2 - (\sum Y_i)^2/n$

$$= 2284.1102 - (235.60)^2/25$$

$$= 63.82.$$

Table 1.3 The Analysis of Variance Table for the Example.

Source	df	SS	MS	Calculated F Value
Total (corrected)	24	63.82		
Regression (b_1)	1	45.59	45.59	57.52
Residual	23	18.23	$s^2 = 0.7926$	

Note that the entries in this table are not in the same order as those in the corresponding theoretical table above. This makes no difference whatsoever. In computer printouts, for example, the order depends on the way in which the program is written. Careful inspection of analysis of variance tables should always be made and it should not be assumed that any particular order is standard. Our estimate of $\sigma^2_{Y \cdot X}$ is $s^2 = 0.7926$ based on 23 degrees of freedom. The F value will be explained shortly.

1.4. Examining the Regression Equation

Up to this point we have made no assumptions at all that involve probability distributions. A number of specified algebraic calculations have been made and that is all. We now make the basic assumptions that, in the model $Y_i = \beta_0 + \beta_1 X + \epsilon_i$, $i = 1, 2, \ldots, n$,

(1) ϵ_i is a random variable with mean zero and variance σ^2 (unknown), that is, $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$.

(2) ϵ_i and ϵ_j are uncorrelated, $i \neq j$, so that

$$\operatorname{cov}(\epsilon_i, \epsilon_j) = 0.$$

Thus

$$E(Y_i) = \beta_0 + \beta_1 X_i, \qquad V(Y_i) = \sigma^2$$

and Y_i and Y_j, $i \neq j$, are uncorrelated. A further assumption, which is not immediately necessary and will be recalled when used, is that

(3) ϵ_i is a normally distributed random variable, with mean zero and variance σ^2 by (1), that is,

$$\epsilon_i \sim N(0, \sigma^2).$$

Under this additional assumption, ϵ_i, ϵ_j are not only uncorrelated but necessarily independent.

Notes

(1) σ^2 may or may not be equal to $\sigma^2_{Y \cdot X}$, the variance about the regression mentioned earlier. If the postulated model is the true model, then $\sigma^2 = \sigma^2_{Y \cdot X}$. If the postulated model is not the true model, then $\sigma^2 < \sigma^2_{Y \cdot X}$. It follows that s^2, the residual mean square which estimates $\sigma^2_{Y \cdot X}$ in any case, is an estimate of σ^2 if the model is correct but not otherwise. If $\sigma^2_{Y \cdot X} > \sigma^2$ we shall say that the postulated model is incorrect or *suffers from lack of fit*. Ways of deciding this will be discussed later.

(2) There is a tendency for errors that occur in many real situations to be normally distributed due to the Central Limit theorem. If an error term such as ϵ is a sum of errors from several sources, then no matter what the probability distribution of the separate errors may be, their sum ϵ will have a distribution that will tend more and more to the normal distribution as the number of components increases, by the Central Limit theorem. An experimental error in practice may be a composite of a meter error, an error due to a small leak in the system, an error in measuring the amount of catalyst used, and so on. Thus the assumption of normality is not unreasonable in most cases. In any case we shall later check the assumption by examining residuals (see Chapter 3).

We now use these assumptions in examining the regression equation.

Standard Error of the Slope b_1; Confidence Interval for β_1.

We know that
$$\begin{aligned} b_1 &= \sum (X_i - \bar{X})(Y_i - \bar{Y})/\sum (X_i - \bar{X})^2 \\ &= \sum (X_i - \bar{X}) Y_i/\sum (X_i - \bar{X})^2 \end{aligned}$$

(since the other term removed from the numerator is $\sum (X_i - \bar{X})\bar{Y} = \bar{Y}\sum (X_i - \bar{X}) = 0$)

$$= \{(X_1 - \bar{X})Y_1 + \cdots + (X_n - \bar{X})Y_n\}/\sum (X_i - \bar{X})^2.$$

Now the variance of a function

$$F = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

is

$$V(F) = a_1^2 V(Y_1) + a_2^2 V(Y_2) + \cdots + a_n^2 V(Y_n),$$

if the Y_i are pairwise uncorrelated and the a_i are constants; furthermore, if $V(Y_i) = \sigma^2$,

$$\begin{aligned} V(F) &= (a_1^2 + a_2^2 + \cdots + a_n^2)\sigma^2 \\ &= (\sum a_i^2)\sigma^2. \end{aligned}$$

In the expression for b, $a_i = (X_i - \bar{X})/\sum (X_i - \bar{X})^2$, since the X_i can be regarded as constants. Hence after reduction

$$V(b_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}. \tag{1.4.1}$$

(*Note:* An implication of this result is of interest. Suppose that, before any data had been collected, we wished to select the X_i values at which to take observations Y_i and wished to do it in a way that would minimize $V(b_1)$. Then the X_i chosen would have to maximize $\sum (X_i - \bar{X})^2$. The theoretical answer to this problem is that some X_i should be located at each of plus and minus infinity. The practical interpretation of this is that the X_i should be located at the extremes of the X-region in which experimental runs are possible. For example, if we wished to perform four runs, two would be placed at each extreme. This result is sensible and correct when the first-order model being tentatively entertained is *precisely the correct one.* When this is not true, and in practice it never really is true, this result may be quite wrong. In fact it has been shown by G. Box and N. Draper (*Journal of the American Statistical Association*, 54, 622–654, 1959) that if the "region of interest" of the X's is scaled to the interval $(-R, R)$ and if we take $\bar{X} = 0$ and a straight line is to be fitted but some second-order tendency exists in the true model, then, the appropriate value for $\sum (X_i - \bar{X})^2$ is not infinity but a number slightly bigger than $NR/3$, where N is the number of X's to be chosen unless the model is nearly correct or the experimental error is very large. The general moral is that conclusions obtained by minimizing variance error only and assuming the postulated model to be correct are likely to be wrong in many practical design situations.)

The standard error of b_1 is the square root of the variance, that is,

$$\text{s.e.}(b_1) = \frac{\sigma}{\{\sum (X_i - \bar{X})^2\}^{1/2}}$$

or, if σ is unknown and we use the estimate s in its place, assuming the model is correct, the *estimated* standard error of b_1 is given by

$$\text{est. s.e.}(b_1) = \frac{s}{\{\sum (X_i - \bar{X})^2\}^{1/2}}. \tag{1.4.2}$$

If we assume that the variations of the observations about the line are normal, that is, that the errors ϵ_i are all from the same normal distribution, $N(0, \sigma^2)$, it can be shown that we can assign $100(1 - \alpha)\%$ confidence limits for b_1 by calculating

$$b_1 \pm \frac{t(n - 2, 1 - \frac{1}{2}\alpha)s}{\{\sum (X_i - \bar{X})^2\}^{1/2}} \tag{1.4.3}$$

where $t(n - 2, 1 - \frac{1}{2}\alpha)$ is the $(1 - \frac{1}{2}\alpha)$ percentage point of a t-distribution, with $(n - 2)$ degrees of freedom (the number of degrees of freedom on which the estimate s^2 is based).

On the other hand, if a test is appropriate, we can test the null hypothesis that β_1 is equal to β_{10}, where β_{10} is a specified value which could be zero, against the alternative that β_1 is different from β_{10} (usually stated "$H_0: \beta_1 = \beta_{10}$ versus $H_1: \beta_1 \neq \beta_{10}$") by calculating

$$\begin{aligned} t &= \frac{(b_1 - \beta_{10})}{\{\text{est. s.e.}(b_1)\}} \\ &= \frac{(b_1 - \beta_{10})\{\sum (X_i - \bar{X})^2\}^{1/2}}{s} \end{aligned} \tag{1.4.4}$$

and comparing $|t|$ with $t(n - 2, 1 - \frac{1}{2}\alpha)$ from a t-table with $(n - 2)$ degrees of freedom—the number on which s^2 is based. The test will be a two-sided test conducted at the $100(1 - \alpha)\%$ level in this form. Calculations for our example follow.

Example (*continued*).

$$\begin{aligned} V(b_1) &= \sigma^2/\sum (X_i - \bar{X})^2 \\ &= \sigma^2/7154.42 \\ \text{est. } V(b_1) &= s^2/7154.42 \\ &= 0.7926/7154.42 \\ &= 0.00011078 \\ \text{est. s.e.}(b_1) &= \sqrt{\text{est. } V(b_1)} = 0.0105 \end{aligned}$$

Suppose $\alpha = 0.05$, so that $t(23, 0.975) = 2.069$. Then 95% confidence limits for β_1 are $b_1 \pm t(23, 0.975) \cdot s/\{\sum (X_i - \bar{X})^2\}^{1/2}$,

or $$-0.0798 \pm (2.069)(0.0105),$$

providing the interval $$-0.1015 \leq \beta_1 \leq -0.0581.$$

In words, the true value β_1 lies in the interval (-0.1015 to -0.0581), and this statement is made with 95% confidence.

We shall also test the null hypothesis that the true value β_1 is zero, or that there is no relationship between atmospheric temperature and the amount of steam used. As noted above, we write (using $\beta_{10} = 0$),

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$$

and evaluate

$$\begin{aligned} t &= b_1/\text{s.e.}(b_1) \\ &= -0.0798/0.0105 \\ &= -7.60. \end{aligned}$$

Since $|t| = 7.60$ exceeds the appropriate critical value of $t(23, 0.975) = 2.069$, $H_0: \beta_1 = 0$ is rejected. (Actually 7.60 also exceeds $t(23, 0.9995)$; we chose a two-sided 95% level test here however so that the confidence interval and the t-test would both make use of the same probability level. In this case we can effectively make the test by examining the confidence interval to see if it includes zero, as described below.) The data we have seen cause us to reject the idea that a linear relationship between Y and X might not exist.

If it had happened that the observed $|t|$ value had been smaller than the critical value we would have said that we *could not reject* the hypothesis. Note carefully that we do not use the word "accept," since we normally cannot accept a hypothesis. The most we can say is that on the basis of certain observed data we cannot reject it. It may well happen, however, that in another set of data we can find evidence which is contrary to our hypothesis and so reject it.

For example, if we see a man who is poorly dressed we may hypothesize, H_0: "This man is poor." If the man walks to save bus fare or avoids lunch to save lunch money, we have no reason to reject this hypothesis. Further observations of this kind may make us feel H_0 is true, but we still cannot accept it unless we know all the true facts about the man. However, a single observation against H_0, such as finding that the man owns a bank account containing $500,000 will be sufficient to reject the hypothesis.

Once we have the confidence interval for β_1 we do not actually have to compute the $|t|$ value for a particular t-test. It is simplest to examine the confidence interval for β_1 and see if it contains the value β_{10}. If it does, then the hypothesis $\beta_1 = \beta_{10}$ cannot be rejected; if it does not, the hypothesis is rejected. This can be seen from Eq. (1.4.4), for $H_0: \beta_1 = \beta_{10}$ is rejected at the $(1 - \alpha)$ level if $|t| > t(n - 2, 1 - \frac{1}{2}\alpha)$, which implies that

$$|b_1 - \beta_{10}| > t(n - 2, 1 - \tfrac{1}{2}\alpha) \cdot s/\{\sum (X_i - \bar{X})^2\}^{1/2}$$

that is, that β_{10} lies outside the limits Eq. (1.4.3).

Standard Error of the Intercept; Confidence Interval for β_0

A confidence interval for β_0 and a test of whether or not β_0 is equal to some specified value can be constructed in a way similar to that just described for β_1. We can show (details in Section 2.3) that

$$\text{s.e.}(b_0) = \left\{\frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}\right\}^{1/2} \sigma.$$

Thus $100(1 - \alpha)\%$ confidence limits for β_0 are given by

$$b_0 \pm t(n - 2, 1 - \tfrac{1}{2}\alpha)\left\{\frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}\right\}^{1/2} s.$$

A t-test for the null hypothesis $H_0: \beta_0 = \beta_{00}$ against the alternative $H_1: \beta_0 \neq \beta_{00}$, where β_{00} is a specified value, will be rejected at the $(1 - \alpha)$ level if β_{00} falls outside the confidence interval, or not be rejected if β_{00} falls inside, or may be conducted separately by finding the quantity

$$t = (b_0 - \beta_{00})\bigg/\left\{\frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}\right\}^{1/2} s$$

and comparing it with percentage points $t(n - 2, 1 - \frac{1}{2}\alpha)$ since $n - 2$ is the number of degrees of freedom on which s^2, the estimate of σ^2, is based. (*Note:* It is also possible to get a *joint confidence region* for β_0 and β_1 simultaneously by applying the formula (2.6.15).)

Standard Error of $\hat{Y}$

We have shown that the regression equation is

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X})$$

where both $\bar{Y}$ and b_1 are subject to error, which will influence $\hat{Y}$. Now if a_i and c_i are constants, and

$$\begin{aligned} a &= a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n, \\ c &= c_1 Y_1 + c_2 Y_2 + \cdots + c_n Y_n, \end{aligned}$$

2. Test the overall regression equation (more specifically, test $H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$ against H_1: not all $\beta_i = 0$) by treating the mean square ratio

$$\frac{[SS(R \mid b_0)/(p-1)]}{s^2} \qquad (2.6.13)$$

as an $F(p-1, v)$ variate where $v = n - p$

Suppose we decide on a specified risk level α. The fact that the observed mean-square ratio exceeds $F(p-1, v, 1-\alpha)$ means that a "statistically significant" regression has been obtained; in other words, the proportion of the variation observed in the data, which has been accounted for by the equation, is greater than would be expected by chance in $100(1-\alpha)\%$ similar sets of data with the same values of n and X. This does not necessarily mean that the equation is useful for predictive purposes. Unless the range of values predicted by the fitted equation is considerably greater than the size of the random error, prediction will often be of no value even though a "significant" F-value has been obtained, since the equation will be "fitted to the errors" only.

Work by J. M. Wetz (in a 1964 Ph.D. thesis, "Criteria for judging adequacy of estimation by an approximating response function," written under the direction of Dr. G. E. P. Box at the University of Wisconsin) suggests that in order that an equation should be regarded as a satisfactory predictor (in the sense that the range of response values predicted by the equation is substantial compared with the standard error of the response), the observed F-ratio of (regression mean square)/(residual mean square) should exceed not merely the selected percentage point of the F-distribution, but about *four times* the selected percentage point. For example, if $p = 11$, $v = 20$, $\alpha = 0.05$, $F(10, 20, 0.95) = 2.35$. Thus the observed F-ratio would have to exceed about 9.4 for the fitted equation to be rated as a satisfactory prediction tool. Since (at the time of writing) work on this topic is not complete, the "four times" rule is given here as a current expedient for assessment of regression equations. It is subject to later confirmation.

3. State that

$$\mathbf{b} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2). \qquad (2.6.14)$$

4. Obtain a joint $100(1-\alpha)\%$ confidence region for *all* the parameters $\boldsymbol{\beta}$ from the equation

$$(\boldsymbol{\beta} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) \leq ps^2F(p, v, 1-\alpha) \qquad (2.6.15)$$

where $F(p, v, 1-\alpha)$ is the $1-\alpha$ point ("upper α-point") of the $F(p, v)$ distribution and where s^2 has the same meaning as in (1) above and the model is assumed correct. In general this will be useful only when p is small as 2, 3, or 4, unless care is taken to present the information in a form in which it can be readily understood. The inequality above provides the equation of an "elliptically shaped" contour in a space which has as many dimensions, p, as there are parameters in $\boldsymbol{\beta}$. We can obtain individual confidence intervals for the various parameters separately from the formula

$$b_i \pm t(v, 1 - \tfrac{1}{2}\alpha) \quad (\text{estimated s.e.}(b_i))$$

where the "estimated s.e.(b_i)" is the square root of the ith diagonal term of the matrix $(\mathbf{X}'\mathbf{X})^{-1}s^2$. (For a calculation of this type when there are two parameters β_0 and β_1, see Eq. (2.3.1), and after replacement of σ^2 by s^2, see pp. 19 and 21. Separate confidence intervals of this type appear in our printouts and are often useful. We de-emphasize them, however, for the following reason. Figure 2.1 illustrates a possible situation that may arise when two parameters are considered. The joint 95% confidence region for the true parameters, β_1 and β_2, is shown as a long thin ellipse and encloses values (β_1, β_2) which the data regard as *jointly* reasonable for the parameters. It takes into account the correlation between the estimates b_1 and b_2. The individual 95% confidence intervals for β_1 and β_2 separately are appropriate for specifying ranges for the individual parameters irrespective of the value of the other parameter. If an attempt is made to interpret these intervals simultaneously—that is (wrongly), regard the rectangle which they define as a joint confidence region—then, for example, it may be thought that the coordinates of the point E provide

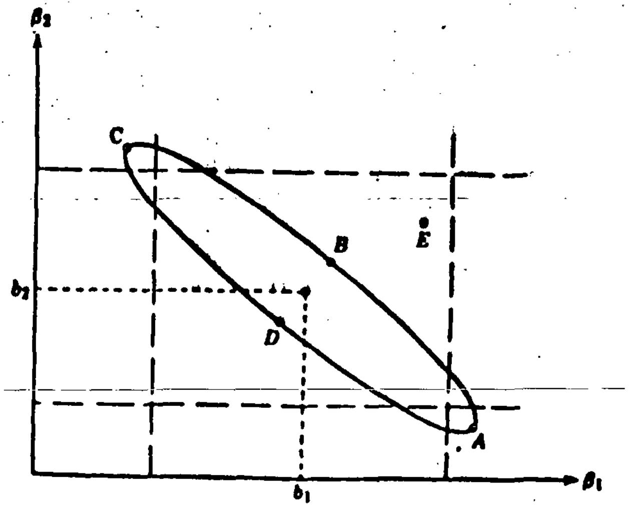


Figure 2.1

CHAPTER 4

TWO INDEPENDENT VARIABLES

4.0. Introduction

Up to this point we have considered, in detail, the first-order linear regression model in one variable X,

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

and shown how the straightforward analysis can be expressed neatly in matrix terms. Usually more complex linear models are needed in practical situations. There are many problems in which a knowledge of more than one independent (or "predictor") variable is necessary in order to obtain better understanding and/or better prediction of a particular response. The matrix approach given at the end of Chapter 2 provides us with a general procedure for extending Chapter 1 results to more complicated linear models. In this chapter, we shall apply the matrix analysis to the first-order linear model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

We shall continue with the example used in Chapter 1 (the data for which are given in Appendix A) and will now add variable number 6 to the problem. So that we are clear about which variables are being considered in the model we shall use the original variable subscripts. Thus our model will be written

$$Y = \beta_0 X_0 + \beta_8 X_8 + \beta_6 X_6 + \epsilon \qquad (4.0.1)$$

where Y = response or number of pounds of steam used per month,
X_0 = dummy variable, whose value is always unity,
X_8 = average atmospheric temperature in the month (in °F), and
X_6 = number of operating days in the month.

The following matrices can then be constructed. (The complete figures for the vector **Y** and the second and third columns of matrix **X** appear in Appendix *A* and are also given on page 116.)

$$\mathbf{Y} = \begin{bmatrix} 10.98 \\ 11.13 \\ 12.51 \\ 8.4 \\ \cdot \\ \cdot \\ \cdot \\ 10.36 \\ 11.08 \end{bmatrix} \quad \mathbf{X} = \begin{array}{c} \begin{matrix} X_0 & X_8 & X_6 \end{matrix} \\ \begin{bmatrix} 1 & 35.3 & 20 \\ 1 & 29.7 & 20 \\ 1 & 30.8 & 23 \\ 1 & 58.8 & 20 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 33.4 & 20 \\ 1 & 28.6 & 22 \end{bmatrix} \end{array} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_8 \\ \beta_6 \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_{24} \\ \epsilon_{25} \end{bmatrix}$$

where **Y** is a (25 × 1) vector,
X is a (25 × 3) matrix,
β is a (3 × 1) vector, and
ε is a (25 × 1) vector.

Using the results of Chapter 2, the least-squares estimates of β_0, β_8, and β_6 are given by

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

where **b** is the vector of estimates of the elements of **β**, provided that **X'X** is nonsingular. Thus

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix} = \left\{ \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 35.3 & 29.7 & 30.8 & \cdots & 28.6 \\ 20 & 20 & 23 & \cdots & 22 \end{bmatrix} \begin{bmatrix} 1 & 35.3 & 20 \\ 1 & 29.7 & 20 \\ 1 & 30.8 & 23 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 28.6 & 22 \end{bmatrix} \right\}^{-1}$$

$$\times \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 35.3 & 29.7 & 30.8 & \cdots & 28.6 \\ 20 & 20 & 23 & \cdots & 22 \end{bmatrix} \begin{bmatrix} 10.98 \\ 11.13 \\ 12.51 \\ \cdot \\ \cdot \\ \cdot \\ 11.08 \end{bmatrix}$$

Note the sizes of the matrices in the above statement:

$$[3 \times 1] = \{[3 \times 25][25 \times 3]\}^{-1}[3 \times 25][25 \times 1].$$

Multiplying the matrices within the large braces, we have

$$\begin{array}{ccc} [3 \times 1] & & [3 \times 3]^{-1} \\ \begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix} & = & \begin{bmatrix} 25.00 & 1315.00 & 506.00 \\ 1315.00 & 76323.42 & 26353.30 \\ 506.00 & 26353.30 & 10460.00 \end{bmatrix}^{-1} \end{array}$$

$$\begin{array}{cc} [3 \times 25] & [25 \times 1] \\ \times \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 35.3 & 29.7 & \cdots & 28.6 \\ 20 & 20 & \cdots & 22 \end{bmatrix} & \begin{bmatrix} 10.98 \\ 11.13 \\ 12.51 \\ \cdot \\ \cdot \\ \cdot \\ 11.08 \end{bmatrix} \end{array}$$

Then,

$$\begin{array}{ccc} [3 \times 1] & [3 \times 3]^{-1} & [3 \times 1] \\ \begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix} = & \begin{bmatrix} 25.00 & 1315.00 & 506.00 \\ 1315.00 & 76323.42 & 26353.30 \\ 506.00 & 26353.30 & 10460.00 \end{bmatrix}^{-1} & \begin{bmatrix} 235.6000 \\ 11821.4320 \\ 4831.8600 \end{bmatrix} \end{array}$$

Next, the inverse of the [3 × 3] matrix is obtained to give

$$\begin{array}{cc} [3 \times 1] & [3 \times 3] \\ \begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix} = & \begin{bmatrix} 2.778747 & -0.011242 & -0.106098 \\ & 0.146207 \times 10^{-3} & 0.175467 \times 10^{-3} \\ \text{(Symmetric)} & & 0.478599 \times 10^{-2} \end{bmatrix} \end{array}$$

$$\begin{array}{c} [3 \times 1] \\ \times \begin{bmatrix} 235.6000 \\ 11821.4320 \\ 4831.8600 \end{bmatrix} \end{array}$$

6 7 1 2 7 9 0

The inverse calculation can be checked by multiplying $(X'X)^{-1}$ by the original $(X'X)$ to give a 3 × 3 unit matrix. Notice that, since the inverse (like the original matrix) is symmetric, only an upper triangular portion of it is recorded. Performing the matrix multiplication gives

$$\begin{array}{ccc} [3 \times 1] & & [3 \times 1] \\ \begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix} & = & \begin{bmatrix} 9.1266 \\ -0.0724 \\ 0.2029 \end{bmatrix} \end{array}$$

Thus, the fitted least squares equation is

$$\hat{Y} = 9.1266 - 0.0724X_8 + 0.2029X_6.$$

Actually, when these matrix calculations are performed by a computer routine, they are not carried through in precisely this way. One reason for this is that large rounding errors may occur when this sequence is followed. This point will be discussed in Section 5.4.

For the record, the algebraic form of the normal equations for the case of two independent variables is as follows:

$$b_0 n + b_1 \sum_{i=1}^{n} X_{1i} + b_2 \sum_{i=1}^{n} X_{2i} = \sum_{i=1}^{n} Y_i$$

$$b_0 \sum_{i=1}^{n} X_{1i} + b_1 \sum_{i=1}^{n} X_{1i}^2 + b_2 \sum_{i=1}^{n} X_{1i}X_{2i} = \sum_{i=1}^{n} X_{1i}Y_i$$

$$b_0 \sum_{i=1}^{n} X_{2i} + b_1 \sum_{i=1}^{n} X_{2i}X_{1i} + b_2 \sum_{i=1}^{n} X_{2i}^2 = \sum_{i=1}^{n} X_{2i}Y_i.$$

We obtained the fitted equation above by a single regression calculation. Actually it is possible to obtain the same equation through a series of simple straight-line regressions. Although this is not the best practical way of obtaining the final equation, it is instructive to consider how it is done. Thus, before we examine the fitted equation in Section 4.2, we shall discuss this alternative procedure.

4.1. Multiple Regression with Two Independent Variables as a Sequence of Straight-Line Regressions

In the previous section, we used least squares to determine the fitted equation

$$\hat{Y} = 9.1266 - 0.0724X_8 + 0.2029X_6.$$

Another way of obtaining this solution is as follows:

1. Plot Y (amount of steam) against X_8 (average atmospheric temperature). This plot is shown in Figure 1.4 in Chapter 1. Note the downward trend; this is reasonable, since as the temperature rises, the need for steam should diminish.

2. Regress Y on X_8. This straight line regression was performed in Chapter 1, and the resulting equation was

$$\hat{Y} = 13.6215 - 0.0798X_8$$

This fitted equation does not predict Y exactly (Table 1.2). Adding a new variable, say X_6 (the number of operating days), to the prediction equation might improve the prediction significantly.

In order to accomplish this, we desire to relate the number of operating days to the amount of unexplained variation in the data after the atmospheric temperature effect has been removed. However, if the atmospheric temperature variations are in any way related to the variability shown in the number of operating days, we must correct for this first. Thus, what we need to do is to determine the relationship between the unexplained variation in the amount of steam used after the effect of atmospheric temperature has been removed, and the remaining variation in the number of operating days after the effect of atmospheric temperature has been removed from it.

3. Regress X_6 on X_8; calculate residuals $X_{6i} - \hat{X}_{6i}$, $i = 1, 2, \ldots, n$. A plot of X_6 against X_8 is shown in Figure 4.1. Using the notation and methods of Chapter 2, the estimates of the regression coefficients are given by

$$\begin{bmatrix} b_0 \\ b_8 \end{bmatrix} = \left\{ \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 35.3 & 29.7 & \cdots & 28.6 \end{bmatrix} \begin{bmatrix} 1 & 35.3 \\ 1 & 29.7 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 28.6 \end{bmatrix} \right\}^{-1}$$

$$\times \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 35.3 & 29.7 & \cdots & 28.6 \end{bmatrix} \begin{bmatrix} 20 \\ 20 \\ \cdot \\ \cdot \\ \cdot \\ 22 \end{bmatrix}$$

0542150

Thus, $\hat{X}_6 = 22.1685 - 0.0367X_8$, and the residuals are shown in Table 4.1.

Table 4.1 Residuals: $X_{6i} - \hat{X}_{6i}$

Observation Number i	X_{6i}	$\hat{X}_{6i}$	$\hat{X}_{6i} - X_{6i}$
1	20	20.87	−0.87
2	20	21.08	−1.08
3	23	21.04	1.96
4	20	20.01	−0.01
5	21	19.92	1.08
6	22	19.55	2.45
7	11	19.44	−8.44
8	23	19.36	3.64
9	21	19.58	1.42
10	20	20.06	−0.06
11	20	20.47	−0.47
12	21	21.11	−0.11
13	21	21.14	−0.14
14	19	20.73	−1.73
15	23	20.45	2.55
16	20	20.39	−0.39
17	22	19.99	2.01
18	22	19.60	2.40
19	11	19.60	−8.60
20	23	19.44	3.56
21	20	19.53	0.47
22	21	20.04	0.96
23	20	20.53	−0.53
24	20	20.94	−0.94
25	22	21.12	0.88

We note that there are two residuals −8.44 and −8.60 which have absolute values considerably greater than the other residuals. They arise from months in which the number of operating days was unusually small, eleven in each case. We can, of course, take the attitude that these are "outliers" and that months with so few operating days should not even be considered in the analysis. However, if we wish to obtain a satisfactory prediction equation which will be valid for *all* months, irrespective of the number of operating days, then it is important to take account of these particular results and develop an equation which makes use of the information they contain. As can be seen from the data and from Figure 4.1 and Table 4.2, if these particular months were ignored, the apparent effect of the number of operating days on the response would be small. This would *not* be because the variable did not affect the response but because the variation actually observed in the variable

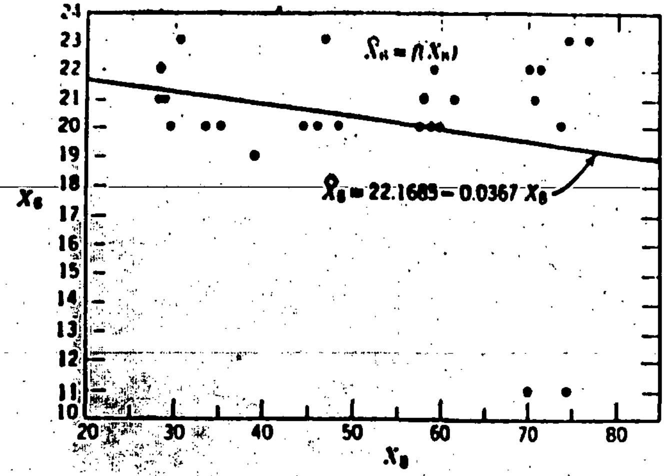


Figure 4.1

was so slight that the variable could not exert any appreciable effect on the response. If a variable appears to have a significant effect on the response in one analysis but not in a second, it may well be that it varied over a wider range in the first set of data than in the second. This, incidentally, is one of the drawbacks of using plant data "as it comes." Quite often the normal operating range of a variable is so slight that no effect on response is revealed, even when the variable does, over larger ranges of operation, have an appreciable effect. Thus designed experiments, which assign levels wider than normal operating ranges, often reveal effects which had not been previously noticed.

4. We now regress $Y - \hat{Y}$ against $X_6 - \hat{X}_6$ by fitting the model

$$(Y_i - \hat{Y}_i) = \beta(X_{6i} - \hat{X}_{6i}) + \epsilon_i.$$

Note that no "β_0" term is required in this first-order model since we are using two sets of residuals whose sums are zero, and thus the line must pass through the origin. (If we did put a β_0 term in, we should find $b_0 = 0$, in any case.) For convenience the two sets of residuals used as data are extracted from Tables 1.2 and 4.1 and are given in Table 4.2. A plot of these residuals is shown in Figure 4.2.

Using a result of Chapter 1,

$$b = \frac{\sum(Y_i - \hat{Y}_i)(X_{6i} - \hat{X}_{6i})}{\sum(X_{6i} - \hat{X}_{6i})^2} = \frac{42.0821}{208.8523} = 0.2015 \subseteq 0$$

Table 4.2 Deviations of $\hat{Y}_i = f(\hat{X}_8)$ and $\hat{X}_{6i} = f(X_8)$ from Y_i and X_{6i}, Respectively

Observation Number i	$Y_i - \hat{Y}_i$	$X_{6i} - \hat{X}_{6i}$
1	0.17	−0.87
2	0.12	−1.08
3	1.34	1.96
4	−0.53	−0.01
5	0.55	1.08
6	−0.80	2.45
7	−1.32	−8.44
8	1.00	3.64
9	−0.16	1.42
10	0.11	−0.06
11	−1.68	−0.47
12	0.87	−0.11
13	0.50	−0.14
14	−0.93	−1.73
15	1.05	2.55
16	−0.17	−0.39
17	1.20	2.01
18	0.07	2.40
19	−1.21	−8.60
20	1.20	3.56
21	−0.19	0.47
22	−0.51	0.96
23	−1.20	−0.53
24	−0.60	−0.94
25	−0.26	0.88

Thus the equation of the fitted line is

$$\widehat{(Y - \hat{Y})} = 0.2015(X_6 - \hat{X}_6).$$

Within the parentheses we can substitute for $\hat{Y}$ and $\hat{X}_6$ as functions of X_8, and the large caret on the left-hand side can then be attached to Y to represent the overall fitted value $\hat{Y} = \hat{Y}(X_6, X_8)$ as follows:

$$[\hat{Y} - (13.6215 - 0.0798X_8)] = 0.2015[X_6 - (22.1685 - 0.0367X_8)]$$

or

$$\hat{Y} = 9.1545 - 0.0724X_8 + 0.2015X_6.$$

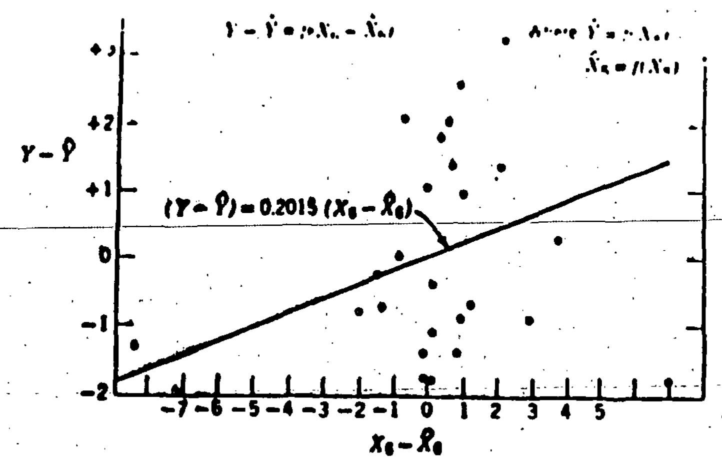


Figure 4.2

The previous result was

$$\hat{Y} = 9.1266 - 0.0724X_8 + 0.2029X_6.$$

In theory these two results are identical; practically, as we can see, slight discrepancies have occurred due to rounding errors. Ignoring rounding errors for the moment we shall show, geometrically, through a simple example, why the two methods should provide us with identical results. (The rest of this section could be omitted at first reading, if desired.)

Consider an example in which we have $n = 3$ observations of the response Y, namely Y_1, Y_2, and Y_3 taken at the three sets of conditions (X_1, Z_1), (X_2, Z_2), (X_3, Z_3). We can plot in three dimensions on axes labeled 1, 2, and 3, with origin at 0, the points $Y \equiv (Y_1, Y_2, Y_3)$, $X \equiv (X_1, X_2, X_3)$, and $Z \equiv (Z_1, Z_2, Z_3)$. The geometrical interpretation of regression is as follows. To regress Y on X we drop a perpendicular YP onto OX. The coordinates of the point P are the fitted values $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}_3$. The length OP^2 is the sum of squares due to the regression, OY^2 is the total sum of squares, and YP^2 is the residual sum of squares. By Pythagoras, $OP^2 + YP^2 = OY^2$, which provides the analysis of variance breakup of the sums of squares (see Figure 4.3).

If we complete the parallelogram which has OY as diagonal and OP and PY as sides, we obtain the parallelogram $OP'YP$ as shown. Then the coordinates of P' are the values of the residuals from the regression of variable Y on variable X. In vector terms we could write

$$\vec{OP} + \vec{OP'} = \vec{OY},$$

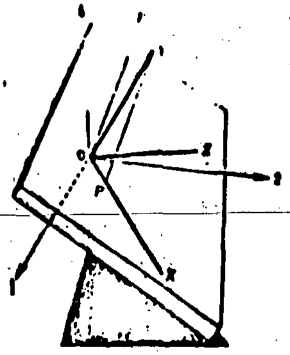


Figure 4.3

or, in "statistical" vector notation,

$$\hat{\mathbf{Y}} + (\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}.$$

This result is true in general for n dimensions. (The only reason we take $n = 3$ is so we can provide a diagram.)

Suppose we wish to regress variable Y on variables X and Z simultaneously. The lines OX and OZ define a plane in three dimensions. We drop a perpendicular YT onto this plane. Then the coordinates of the point T are the fitted values $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}_3$ for *this* regression. OT^2 is the regression sum of squares, YT^2 is the residual sum of squares, and OY^2 is the total sum of squares. Again, by Pythagoras, $OY^2 = OT^2 + YT^2$ which, again, gives the sum of squares breakup we see in the analysis of variance table. Completion of the parallelogram $OT'YT$ with diagonal OY and sides OT, TY provides OT', the vector of residuals of this regression, and the coordinates of T' give the residuals $\{(Y_1 - \hat{Y}_1), (Y_2 - \hat{Y}_2), (Y_3 - \hat{Y}_3)\}$ of the regression of Y on X and Z simultaneously. Again, in vector notation,

$$\vec{OT} + \vec{OT'} = \vec{OY}$$

or, in "statistical" vector notation,

$$\hat{\mathbf{Y}} + (\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}$$

for this regression (see Figure 4.4).

As we saw in the numerical example above, the same final residuals should arise (ignoring rounding) if we do the regressions (1) Y on X, and

(2) Z on X, and then regress the residuals of (1) on the residuals of (2). That this is true can be seen geometrically as follows. Figure 4.5 shows three parallelograms in three dimensional space.

1. $OP'YP$ from the regression of Y on X,
2. $OQ'ZQ$ from the regression of Z on X, and
3. $OT'YT$ from the regression of Y on X and Z simultaneously.

Now the regression of the residuals of (1) onto the residuals of (2) is achieved by dropping the perpendicular from P' onto OQ'. Suppose the point of impact is R. Then a line through O parallel to RP' and of length RP' will be the residual vector of the two-step regression of Y on X and Z. However, the points O, Q', Z, P, Q, X, and T all lie in the plane π defined by OZ and OX. Thus so does the point R. Since $OP'YP$ is a parallelogram, and $P'R$ and YT are perpendicular to plane π, $P'R = YT$ in length. Since $TY = OT'$, it follows that $OT' = RP'$. But OT', RP', and TY are all parallel and perpendicular to plane π. Hence $OT'P'R$ is a parallelogram from which it follows that $\overrightarrow{OT'}$ is the vector of residuals from the two-step regression. Since it originally resulted from the regression of Y on Z and X together the two methods must be equivalent. Thus we can see that the planar regression of Y on X and Z together can be regarded as the totality of successive straight-line regressions of

1. Y on X,
2. Z on X, and
3. residuals of (1) on the residuals of (2).

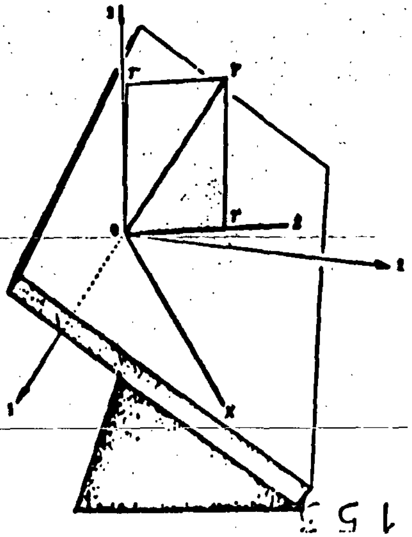


Figure 4.4

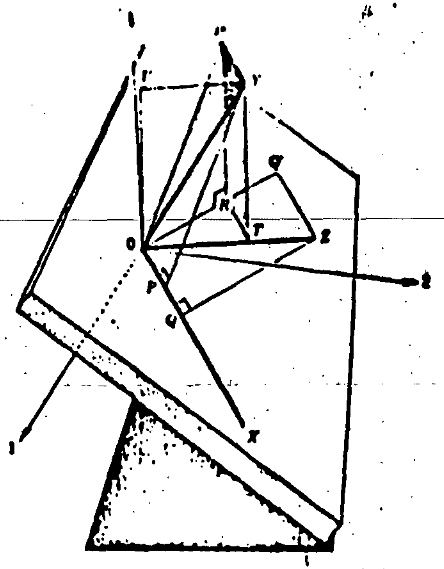


Figure 4.5

The same result is obtained if the roles of Z and X are interchanged. All linear regressions can be broken down into a series of simple regressions in this way. (For an application, see page 180.)

4.2. Examining the Regression Equation

How Useful Is the Equation, $Y = f(X_8, X_6)$?

Utilizing the work given in Chapters 1 and 2, we shall now consider the equation obtained for $\hat{Y}$ as a function of X_8 and X_6. We can calculate the residuals, using the fitted equation and the observed points. These residuals are shown in Table 4.3. The regression analysis of variance is as follows:

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Total (uncorrected)	25	2284.1102		
Mean (b_0)	1	2220.2944		
Total (corrected)	24	63.8158		
Regression | b_0	2	54.1871	27.0936	61.8999
Residual	22	9.6287	0.4377	

On the basis of an α risk of 0.05, the least squares equation

$$\hat{Y} = 9.1266 - 0.0724X_8 + 0.2029X_6$$

is a good predictor; the calculated $F = 61.8999$ for regression is greater than the tabulated $F(2, 22, 0.95) = 3.44$.

Table 4.3

Obs. No.	X_8	X_6	Y	$\hat{Y}$	Residual
1	35.3	20	10.98	10.63	0.35
2	29.7	20	11.13	11.03	0.10
3	30.8	23	12.51	11.56	0.95
4	58.8	20	8.40	8.93	−0.53
5	61.4	21	9.27	8.94	0.33
6	71.3	22	8.73	8.43	0.30
7	74.4	11	6.36	5.97	0.39
8	76.7	23	8.50	8.24	0.26
9	70.7	21	7.82	8.27	−0.45
10	57.5	20	9.14	9.02	0.12
11	46.4	20	8.24	9.82	−1.58
12	28.9	21	12.19	11.29	0.90
13	28.1	21	11.88	11.35	0.53
14	39.1	19	9.57	10.15	−0.58
15	46.8	23	10.94	10.40	0.54
16	48.5	20	9.58	9.67	−0.09
17	59.3	22	10.09	9.30	0.79
18	70.0	22	8.11	8.52	−0.41
19	70.0	11	6.83	6.29	0.54
20	74.5	23	8.88	8.40	0.48
21	72.1	20	7.68	7.96	−0.28
22	58.1	21	8.47	9.18	−0.71
23	44.6	20	8.86	9.96	−1.10
24	33.4	20	10.36	10.77	−0.41
25	28.6	22	11.08	11.52	−0.44
			235.60		$\Sigma(Y_i - \hat{Y}_i) = 0$
			$\bar{Y} = 9.424$		$\Sigma(Y_i - \hat{Y}_i)^2 = 9.6432$

A graph of the observed values of Y and the fitted $\hat{Y}$'s is shown in Figure 4.6. The graph indicates that the fitted model is a good predictor of monthly steam usage. However, has the addition of X_6 to the model been useful?

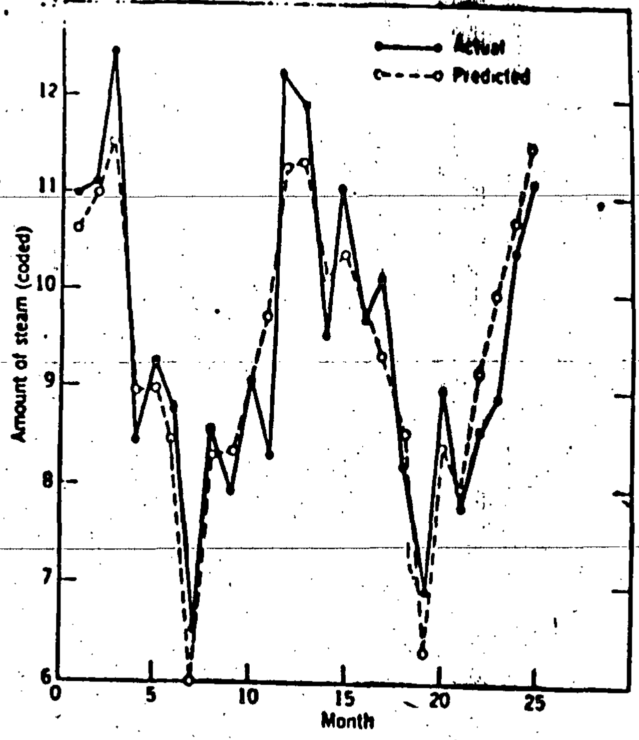


Figure 4.6 Amount of steam used in a plant by month.

What Has Been Accomplished by the Addition of a Second Independent Variable (Namely X_6)?

• There are several useful criteria which can be applied to answer this question, and we now discuss them.

THE SQUARE OF THE MULTIPLE CORRELATION COEFFICIENT, R^2. The square of the multiple correlation coefficient R^2 is defined as (see Eq. 2.6.11)

$$R^2 = \frac{\text{Sum of squares due to regression} \mid b_0}{\text{Total (corrected) sum of squares}}.$$

It is often stated as a percentage, $100R^2$. The larger it is, the better the fitted equation explains the variation in the data. We can compare the value of R^2 at each stage of the regression problem:

STEP 1. $Y = f(X_8)$.

Regression equation	$100R^2$
$\hat{Y} = 13.6215 - 0.0798X_8$	71.44%

STEP 2. $Y = f(X_8, X_6)$.

Regression equation	$100R^2$
$\hat{Y} = 9.1266 - 0.0724X_8 + 0.2029X_6$	84.89%

Thus, we see a substantial increase in R^2. However, this statistic must be used with caution, since one can always make $R^2 = 1$ as described on page 63.

In addition, given that the number of observations is much greater than the number of potential X variables under consideration, the addition of a new variable will always increase R^2 but it will not necessarily increase the precision of the estimate of the response. This is because the reduction in the residual sum of squares may be less than the original residual mean square. Since one degree of freedom is removed from the residual degrees of freedom as well, the resulting mean square may get larger. An example of this can be seen in Appendix B (pp. 387, 395) which we have not yet discussed. We can make the following comparison:

			Residual		
Page	R^2	Variables in Regression Model	Sum of Squares	Degrees of Freedom	Mean Square
387	98.23	1, 2, 3	48.11	9	5.35
395	98.24	1, 2, 3, 4	47.86	8	5.98

We see that although an extra variable has been included in the regression model, the residual mean square has increased since the extra variable produced a residual sum of squares reduction of $48.11 - 47.86 = 0.25 < 5.35$ for the loss of one degree of freedom. The value of R^2 has increased slightly, however.

THE STANDARD ERROR OF ESTIMATE, s. The residual mean square s^2 is an estimate of $\sigma^2_{Y\cdot X}$, the variance about the regression. Before and after adding a variable to the model, we can check

$$s = \sqrt{\text{residual mean square}}.$$

Examination of this statistic indicates that the smaller it is the better that is, the more precise will be the predictions. Since s can be made zero by including enough parameters in the model—just as R^2 can be made unity—this criterion must also be used cautiously. Provided there are few repeats and many degrees of freedom for error remaining, reduction of s is desirable. In our example at Step 1,

$$s = \sqrt{0.7926} = 0.89.$$

At Step 2,

$$s = \sqrt{0.4377} = 0.66.$$

Thus, the addition of X_6 has decreased s and improved the precision of estimation.

THE STANDARD ERROR OF ESTIMATE s, AS A PERCENTAGE OF THE MEAN RESPONSE. Another way of looking at the decrease in s is to consider it in relation to the response. In our example, at Step 1, s as a percentage of mean $\bar{Y}$ is

$$0.89/9.424 = 9.44\%.$$

At Step 2, s as a percentage of mean $\bar{Y}$ is

$$0.66/9.424 = 7.00\%.$$

Thus, the addition of X_6 has reduced the standard error of estimate down to about 7% of the mean response. Whether this level of precision is satisfactory or not is a matter for the experimenter to decide, on the basis of his prior knowledge and personal feelings.

THE SEQUENTIAL F-TEST CRITERION (SHOWING THE ADDITIONAL CONTRIBUTION OF X_6 GIVEN THAT X_8 IS ALREADY IN THE EQUATION). This method of assessing the value of X_6 as a variable added to $Y = f(X_8)$ consists in breaking down the sum of squares due to regression into two parts as follows:

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Total (uncorrected)	25	2284.1102		
Mean (b_0)	1	2220.2944		
Total (corrected)	24	63.8158		
Regression $\mid b_0$	2	54.1871	27.0936	61.8999
due to $b_8 \mid b_0$	1	45.5924	45.5924	104.1636
due to $b_6 \mid b_8, b_0$	1	8.5947	8.5947	19.6361
Residual	22	9.6287	0.4377	

Since 19.5761 exceeds $F(1, 22, 0.95) = 4.30$, the addition of the variable X_6 has been worthwhile. This F-test is usually called the "sequential F-test" (see section 2.9).

THE PARTIAL F TEST CRITERION (see Section 2.9). Another way of assessing the value of X_6 is to consider the order of the two variables in the least squares procedure. For example, the following questions could be asked:

1. If we had put X_6 into the equation first what would its contribution have been?
2. Given that X_6 was used first, what contribution does X_8 make when added to regression?

These questions are answered by performing the calculations shown above, but in reverse order. The results are as follows:

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Total (uncorrected)	25	2284.1102		
Mean (b_0)	1	2220.2944		
Total (corrected)	24	63.8158		
Regression $\mid b_0$	2	54.1871	27.0936	61.8999
due to $b_6 \mid b_0$	1	18.3424	18.3424	41.9063
due to $b_8 \mid b_0, b_6$	1	35.8447	35.8447	81.8933
Residual	22	9.6287	0.4377	

Note that the contribution of X_8 above is more important than is its contribution after X_6 has been introduced. Note also that this is reflected in the observed value of F for X_8 in the two steps; that is,

Step 1 104.1636,

Step 2 81.8933.

However, X_8 is still the more important variable in both cases, since its contribution in reducing the residual sum of squares is the larger, regardless of the order of introduction of the variables.

Standard Error of b_i

Using the result given in Section 2.6, the variance-covariance matrix of **b** is $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$.

Thus, variance of $b_i = V(b_i) = c_{ii}\sigma^2$, where c_{ii} is the diagonal element in $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to the ith variable.

The covariance of b_i, $b_j = c_{ij}\sigma^2$, where c_{ij} is the off-diagonal element in $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to the intersection of the ith row and jth column, or jth row and ith column, since $(\mathbf{X}'\mathbf{X})^{-1}$ is symmetric.

Thus the s.e. of b_i is $\sigma\sqrt{c_{ii}}$. For example, using figures from pages 106 and 120 the estimated standard error of b_8 is obtained as follows:

$$\begin{aligned}\text{est. var } b_8 &= s^2c_{88}\\ &= (0.4377)(0.146207 \times 10^{-3})\\ &= 0.639948 \times 10^{-4}.\end{aligned}$$

Thus est. s.e. $b_8 = \sqrt{\text{var } b_8} = \sqrt{0.639948 \times 10^{-4}} = 0.008000.$

Confidence Limits for the True Mean Value of Y, Given a Specific Set of X's

The predicted value $\hat{Y} = b_0 + b_1X_1 + \cdots + b_pX_p$ is an estimate of

$$E(Y) = \beta_0 + \beta_1X_1 + \cdots + \beta_pX_p.$$

The variance of $\hat{Y}$, $V[b_0 + b_1X_1 + \cdots + b_pX_p]$ is

$$V(b_0) + X_1^2V(b_1) + \cdots + X_p^2V(b_p) + 2X_1 \operatorname{covar}(b_0, b_1) + \cdots + 2X_{p-1}X_p \operatorname{covar}(b_{p-1}, b_p).$$

This expression can be written very conveniently in matrix notation as follows, where $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$.

$$\begin{aligned}V(\hat{Y}) &= \sigma^2(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)\\ &= \sigma^2[1 \quad X_1 \quad \cdots \quad X_p]\begin{bmatrix} c_{00} & c_{01} & \cdots & c_{0p}\\ c_{10} & c_{11} & \cdots & c_{1p}\\ \cdot & \cdot & & \cdot\\ \cdot & \cdot & & \cdot\\ \cdot & & \cdot & \cdot\\ c_{p1} & & & c_{pp}\end{bmatrix}\begin{bmatrix}1\\ X_1\\ \cdot\\ \cdot\\ \cdot\\ X_p\end{bmatrix}\end{aligned}$$

Thus, the $1 - \alpha$ confidence limits on the true mean value of Y at $\mathbf{X}_0$ are given by

$$\hat{Y} \pm t((n - p - 1), 1 - \tfrac{1}{2}\alpha) \cdot s\sqrt{\mathbf{X}_0'\mathbf{C}\mathbf{X}_0}.$$

For example, the variance of $\hat{Y}$ for the point in the X space ($X_8 = 32$, $X_6 = 22$) is obtained as follows:

$$\begin{aligned}\operatorname{var}(\hat{Y}) &= s^2(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)\\ &= (0.4377)(1, 32, 22)\\ &\quad \times \begin{bmatrix} 2.778747 & -0.011242 & -0.106098\\ -0.011242 & 0.146207 \times 10^{-3} & 0.175467 \times 10^{-3}\\ -0.106098 & 0.175467 \times 10^{-3} & 0.478599 \times 10^{-3}\end{bmatrix}\begin{bmatrix}1\\ 32\\ 22\end{bmatrix}\\ &= (0.4377)(0.104140) = 0.045582.\end{aligned}$$

The 95% confidence limits on the true mean value of Y at $X_8 = 32$, $X_6 = 22$ are given by

$$\begin{aligned}\hat{Y} \pm t(22, 0.975) \cdot s\sqrt{\mathbf{X}_0'\mathbf{C}\mathbf{X}_0} &= 11.2736 \pm (2.074)(0.213499)\\ &= 11.2736 \pm 0.4418\\ &= 10.8318, 11.7154\end{aligned}$$

These limits are interpreted as follows. Suppose repeated samples of Y's are taken of the same size each time and at the same fixed values of (X_8, X_6) as were used to determine the fitted equation obtained above. Then of all the 95% confidence intervals constructed for the mean value of Y for $X_8 = 32$, $X_6 = 22$, 95% of these intervals will contain the true mean value of Y at $X_8 = 32$, $X_6 = 22$. From a practical point of view we can say that there is a 0.95 probability that the statement, the true mean value of Y at $X_8 = 32$, $X_6 = 22$ lies between 10.8318 and 11.7154, is correct.

Confidence Limits for the Mean of g, Observations Given a Specific Set of X's

These limits are calculated from

$$\hat{Y} \pm t(v, 1 - \tfrac{1}{2}\alpha) \cdot s\sqrt{1/g + \mathbf{X}_0'\mathbf{C}\mathbf{X}_0}.$$

For example, the 95% confidence limits for an individual observation for the point $(X_8 = 32, X_6 = 22)$ are

$$\begin{aligned}&\hat{Y} \pm t(22, 0.975) \cdot s\sqrt{1 + \mathbf{X}_0'\mathbf{C}\mathbf{X}_0}\\ &\qquad = 11.2736 \pm (2.074)(0.661589)\sqrt{1 + 0.10413981}\\ &\qquad = 11.2736 \pm (2.074)(0.661589)(1.050781)\\ &\qquad = 11.2736 \pm 1.4418\\ &\qquad = 9.8318, 12.7154\end{aligned}$$

Examining the Residuals

The residuals shown in Table 4.3, page 116, could be examined to see if they provide any indication that the model is inadequate. We leave this as an exercise except for the following comments:

1. Residual versus $\hat{Y}$ (Figure 4.7). No unusual behavior is indicated by this plot.
2. Residual versus Y (Figure 4.8). There is some evidence that the larger observed values were underpredicted by the model; that is, six out of seven largest values of Y have positive residuals. This means that the model should be amended to provide better prediction at higher stea

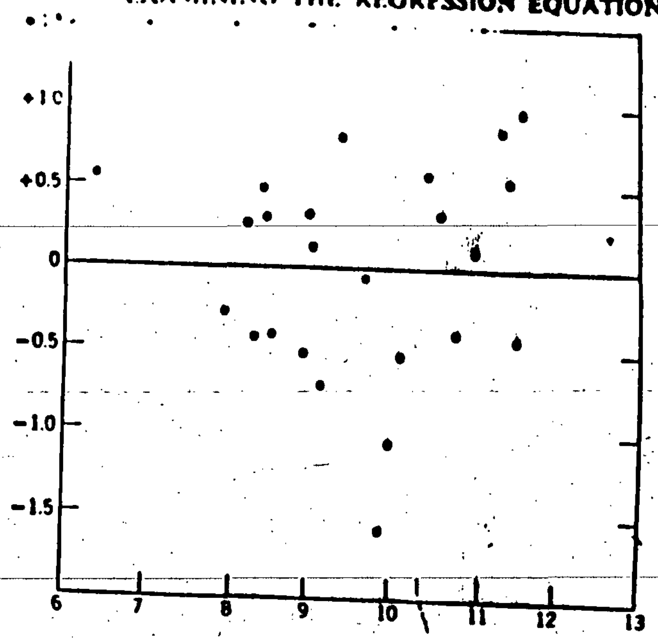


Figure 4.7 Residual versus $\hat{Y}$.

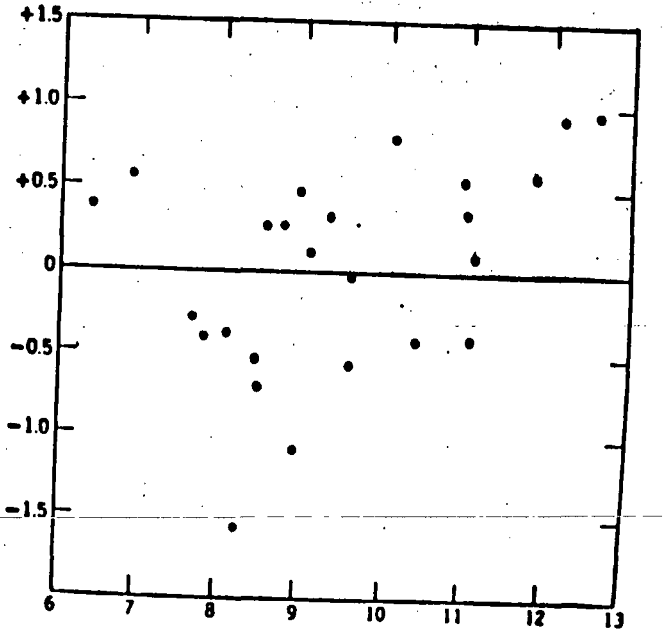


Figure 4.8 Residual versus Y.

levels. Although we do not follow up this point here, additional effort would normally be made to find one or more independent variables which might be added to the model.

3. The runs test indicates no evidence of time dependent nonrandomness.

EXERCISES

A. *Multiple Regression Problem*

Data

X_0	X_1	X_2	Y
1	1	8	6
1	4	2	8
1	9	−8	1
1	11	−10	0
1	3	6	5
1	8	−6	3
1	5	0	2
1	10	−12	−4
1	2	4	10
1	7	−2	−3
1	6	−4	5

Requirements

1. Using least squares procedures, estimate the β's in the model:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

(*Hint:* Use the normal equations for ease of computation.)

2. Write out the Analysis of Variance table.
3. Using $\alpha = 0.05$, test to determine if the overall regression is statistically significant.
4. Calculate the square of the multiple correlation coefficient, namely R^2. What portion of the total variation is explained by the two variables?
5. The inverse of the X′X matrix for this problem is as follows:

$$\begin{bmatrix} 4.3705 & -0.8495 & -0.4086 \\ -0.8495 & 0.1690 & 0.0822 \\ -0.4086 & 0.0822 & 0.0422 \end{bmatrix}$$

Using the results of the Analysis of Variance table with this matrix, calculate the following:

a. Variance of b_1.
b. Variance of b_2.
c. The variance of the predicted value of Y for the point $X_1 = 3$, $X_2 = 5$.

6. How useful is the regression using X_1 alone? What does X_2 contribute, given that X_1 is already in the regression?
7. How useful is the regression using X_2 alone? What does X_1 contribute, given that X_2 is already in the regression?
8. What are your conclusions?

B. The table below gives twelve sets of observations on three variables X, Y, Z. Find the regression plane of X on Y and Z - that is, the linear combination of Y and Z that best predicts the value of X when only Y and Z are given. By constructing an analysis of variance table for X, or otherwise, test whether it is advantageous to include both Y and Z in the prediction formula.

X	Y	Z
1.52	98	77
1.41	76	139
1.16	58	179
1.45	94	95
1.24	73	142
1.21	57	186
1.63	97	82
1.38	91	100
1.37	79	125
1.36	92	96
1.40	92	99
1.03	54	190

(*Cambridge Diploma*, 1949)

C. The data below are selected from a much larger body of data referring to candidates for the General Certificate of Education who were being considered for a special award. Here, Y denotes the candidate's total mark, out of 1000, in the G.C.E. examination. Of this mark the subjects selected by the candidate account for a maximum of 800; the remainder, with a maximum of 200, is the mark in the compulsory papers—"General" and "Use of English"—this mark is shown as X_1. X_2 denotes the candidate's mark, out of 100, in the compulsory School Certificate English Language paper taken on a previous occasion.

Compute the multiple regression of Y on X_1 and X_2, and make the necessary tests to enable you to comment intelligently on the extent to which current performance in the compulsory papers may be used to predict aggregate performance in the G.C.E. examination, and on whether previous performance in School Certificate English Language has any predictive value independently of what has already emerged from the current performance in the compulsory papers.

Candidate	Y	X_1	X_2	Candidate	Y	X_1	X_2
1	476	111	68	9	645	117	59
2	457	92	46	10	556	94	97
3	540	90	50	11	634	130	57
4	551	107	59	12	637	118	51
5	575	98	50	13	390	91	44
6	698	150	66	14	562	118	61
7	545	118	54	15	560	109	66
8	574	110	51				

(*Cambridge Diploma*, 1953)

054 215

Attachment G

TRIC
METHODS
r and Douglas A.

Series in Probabil-
l Statistics, edited
S. Hunter, D. G.
Watson

ortance of non-
s as a significant
statistics is recog-
ns, modern non-
es are not widely
ists. Nonparamet-
ds makes modern
ods accessible to
the scientists in a
es. Rather than
to nonparametric
provide balanced
d recent advances
and multiple com-

ures that are easy
o understand, the
nditions for valid
parametric tech-
ll learn which non-
es are appropriate
s and will develop
for the motivation
ds.
from many disci-
ronomy, biology,
ion, engineering,
nce, psychology,
e science illustrate
These examples
te applicability of
and nonparametric
r. Extensive tables
menting the proce-
ted.
stical *Methods* is
icians, psycholo-
d sociologists. It will
in the fields of
science, business,
a text, the book is
undergraduate

CATION IN APPLIED STATISTICS

Nonparametric Statistical Methods

MYLES HOLLANDER

The Florida State University

DOUGLAS A. WOLFE

The Ohio State University

JOHN WILEY & SONS New York . London . Sydney . Toronto

structuring procedures. One of the major variables of the study was that of "psychotherapeutic attraction." The basic data in Table 2 consist of the raw scores for this measure according to each of the four experimental conditions. Apply procedure (6), with the correction for ties given by (7).

2. Show directly, or illustrate by means of an example, that the maximum value of H is $H_{\max} = (N^3 - \sum_{j=1}^{k} n_j^3)/(N(N+1))$. For what rank configurations is this maximum achieved?

2. A DISTRIBUTION-FREE TEST FOR ORDERED ALTERNATIVES (JONCKHEERE, TERPSTRA)

Procedure. To test H_0 (2) against alternatives (see Comment 9) of the form

$$H_a: \tau_1 \leq \tau_2 \leq \cdots \leq \tau_k, \tag{8}$$

where at least one of the inequalities is strict,

1. Compute $k(k-1)/2$ Mann-Whitney counts U_{uv}, $u < v$, where

$$U_{uv} = \sum_{i=1}^{n_u} \sum_{i'=1}^{n_v} \phi(X_{iu}, X_{i'v}) \tag{9}$$

and $\phi(a, b) = 1$ if $a < b$, 0 otherwise. That is, U_{uv} is the number of sample u before sample v precedences.

2. Let

$$J = \sum_{u<v}^{k} U_{uv} = \sum_{u=1}^{k-1} \sum_{v=u+1}^{k} U_{uv} \tag{10}$$

be the sum of these $k(k-1)/2$ Mann-Whitney counts.

3. At the α level of significance,

$$\begin{aligned} &\text{reject } H_0 \quad \text{if } J \geq j(\alpha, k, (n_1, \ldots, n_k)), \\ &\text{accept } H_0 \quad \text{if } J < j(\alpha, k, (n_1, \ldots, n_k)), \end{aligned} \tag{11}$$

where the constant $j(\alpha, k, (n_1, \ldots, n_k))$, which satisfies the equation $P_0(J \geq j(\alpha, k, (n_1, \ldots, n_k))) = \alpha$, is obtained from Table A.8.

Large Sample Approximation. Set

$$J^* = \frac{J - E_0(J)}{[\operatorname{var}_0(J)]^{1/2}} = \frac{J - \left\{\left(N^2 - \sum_{j=1}^{k} n_j^2\right)\Big/4\right\}}{\left\{\left[N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)\right]\Big/72\right\}^{1/2}}. \tag{12}$$

When H_0 is true, the statistic J^* has an asymptotic [min $(n_1, \ldots, n_k)$ tending to infinity] $N(0, 1)$ distribution. The approximate α-level test is

$$\begin{aligned} &\text{reject } H_0 \quad \text{if } J^* \geq z_{(\alpha)}, \\ &\text{accept } H_0 \quad \text{if } J^* < z_{(\alpha)}. \end{aligned} \tag{13}$$

Ties. Replace $\phi(a, b)$ by $\phi^*(a, b) = 1$ if $a < b$, ½ if $a = b$, 0 otherwise, so that for each between sample comparison where there is a tie, the contribution to the Mann-Whitney count will be ½.

Example 2. Hundal (1969) described a study designed to assess the purely motivational effects of knowledge of performance in a repetitive industrial task. The task was to grind a metallic piece to a specified size and shape. Eighteen male workers were divided randomly into three groups. The subjects in the control group A received no information about their output, subjects in group B were given a rough estimate of their output, and subjects in group C were given accurate information about their output and could check it further by referring to a figure that was placed before them. The basic data in Table 3 consist of the numbers of pieces processed by each subject in the experimental period.

We apply Jonckheere's test with the notion that a deviation from H_0 is likely to be in the direction of increased output with increased degree of knowledge of performance.

From (9) we obtain

$$U_{12} = 22, \quad U_{13} = 30.5, \quad U_{23} = 26.5,$$

and from (10) we have

$$J = 22 + 30.5 + 26.5 = 79.$$

From Table A.8 we find $j(.0231, 3, (6, 6, 6)) = 79$ and thus using procedure (11) the lowest level at which we can reject H_0 (2) is .0231. Now let us apply the large sample approximation and compare it with the exact test. From (12) we compute

$$J^* = \frac{\{79 - (\frac{1}{4})[(18)^2 - (6^2 + 6^2 + 6^2)]\}}{\{[(18)^2(39) - 3(6)^2(15)]/72\}^{1/2}} = \frac{25}{12.37} = 2.02.$$

Table 3. Number of pieces processed

Control (no information)	Group B (rough information)	Group C (accurate information)
40 (5.5)*	38 (2.5)	48 (18)
35 (1)	40 (5.5)	40 (5.5)
38 (2.5)	47 (17)	45 (15)
43 (10.5)	44 (13)	43 (10.5)
44 (13)	40 (5.5)	46 (16)
41 (8)	42 (9)	44 (13)

Source. P. S. Hundal (1969).

* Although we do not need to perform the joint ranking to compute Jonckheere's statistic, we give these ranks here for later use in Section 3D.

Hence, using the approximate procedure (13), the lowest level at which we reject H_0 is .0217. Thus, for the task considered by Hundal, both the exact test and the large sample approximation indicate strong evidence of increased output with increase in degree of knowledge of performance.

Comments

9. In addition to degrees of knowledge of performance, other examples of ordered treatments are quality of materials, amount of practice, intensity of a stimulus, and temperature. Jonckheere's test should be preferred to the Kruskal-Wallis test (Section 1) when the treatments are ordered and the experimenter expects a deviation from H_0 to be in a particular direction. (If the direction expected is not the natural ordering $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$ of H_a, simply relabel the treatments so that the postulated order agrees with the natural order used here.) Note that the Kruskal-Wallis statistic does not utilize the partial prior information in a postulated alternative ordering. The statistic H (4) takes on the same value for all possible (k!) labelings of the treatments.

10. Consider J (10) and note that the term $\sum_{u<v}^{k} U_{uv}$ takes the postulated ordering into account. Consider, for simplicity, the case $k = 3$. Then $\sum_{u<v}^{3} U_{uv} = U_{12} + U_{13} + U_{23}$, and if $\tau_1 < \tau_2 < \tau_3$, U_{12} would tend to be larger than $n_1n_2/2$ (its null expectation); U_{13} would tend to be larger than $n_1n_3/2$; U_{23} would tend to be larger than $n_2n_3/2$ and, consequently, $J = U_{12} + U_{13} + U_{23}$ would tend to be larger than its null expectation $(n_1n_2 + n_1n_3 + n_2n_3)/2 = \{[N^2 - (n_1^2 + n_2^2 + n_3^2)]/4\}$. This serves as partial motivation for the J test.

11. A little thought will convince the reader that J can be computed from the joint ranking of all $N = \sum_{j=1}^{k} n_j$ observations. That is, although we do not need to perform this joint ranking in order to compute J, given the ranking we can, without knowledge of the actual X_{ij} values, retrieve the value of J. Thus one way to obtain the null distribution of J is to follow the method of Comment 5; namely, use the fact that under H_0 (2) all $N!/(\prod_{j=1}^{k} n_j!)$ rank assignments are equally likely, and compute the associated value of J for each possible ranking. Consider how this would work in the small sample size case of $k = 3$, $n_1 = n_2 = n_3 = 2$, which was used in Comment 5. We can easily calculate the value of J for each of the 15 rank configurations displayed in Comment 5. However, we cannot eliminate the calculations corresponding to the other 75 rank configurations (as we were able to do in the case of the Kruskal-Wallis H statistic), since J does depend on the particular numbering of the treatments. For example, consider rank configurations (c) and (a^*):

(a)

I	II	III
1	3	5
2	4	6

(a^*)

I	II	III
3	1	5
4	2	6.

Note that (a) and (a^*) may be viewed as the same except that we have interchanged the roles of treatments I and II. The J value for (a) is readily found to be $J = 12$, whereas for (a^*) we have $J = 8$. [The statistic H (4) takes on the same value, 4.57, for (a) and (a^*).] Without resorting to complete enumeration, we can check a value in Table A.8 as follows. Configuration (a) will yield the largest possible value of J, namely, $J = 12$, and no other configuration will yield a value that large. Thus $P_0(J \geq 12) = P_0(J = 12) = \frac{1}{90} = .0111$, which agrees with the appropriate entry in Table A.8.

12. Table A.8 gives critical values for $n_1 \leq n_2 \leq n_3$ situations. However, critical points for (n_1, n_2, n_3) configurations not in this order can be obtained by simply putting the three sample sizes in increasing order and then entering Table A.8. (This is a consequence of certain symmetry properties of the distribution of J.) Thus, for example, to find critical values for the case $n_1 = 4$, $n_2 = 6$, $n_3 = 2$, enter Table A.8 at $n_1 = 2$, $n_2 = 4$, $n_3 = 6$.

13. For $k = 2$, the procedure defined by (11) reduces to the one-sided Mann-Whitney-Wilcoxon test (Section 4.1).

Properties

1. Consistency: The condition n_j/N tends to λ_j, $0 < \lambda_j < 1$, $j = 1, \ldots, k$, is sufficient to insure that the test defined by (11) is consistent against the H_a (8) alternatives. For a more general consistency statement, see Terpstra (1952).

2. Efficiency: See Puri (1965) and Section 5.

REFERENCES. The test based on the J statistic was proposed by Terpstra (1952) and independently by Jonckheere (1954a). The first generalization of the Mann-Whitney-Wilcoxon two-sample test, with ordered alternatives in mind, was given by Whitney (1951). Whitney treated the case $k = 3$, and his procedures are not equivalent to the J test when the latter is specialized to $k = 3$. Chacko (1963) proposed a rank analog of a normal theory ordered alternatives test developed by Bartholomew (1959a, 1959b, 1961a, 1961b). Puri (1965) generalized Jonckheere's test to a class of tests including a normal scores analog of the J test. Further generalizations were given by Tryon and Hettmansperger (1971). For a different approach to the utilization of partial prior information, see Abelson and Tukey (1963).

PROBLEMS

3. Apply Jonckheere's test to the data of Table 2 using the postulated ordering $\tau_1 < \tau_2 < \tau_3 < \tau_4$.

4. The statistic J can be computed either from (a) the joint ranking of all $N = \sum_{j=1}^{k} n_j$ observations or from (b) $k(k-1)/2$ "two-sample" rankings. Explain.

Table A.8 (continued)

k = 3

n_1 n_2 n_3	α	$f(\alpha, 3, (n_1, n_2, n_3))$
7 7 7	.5000	74
	.4747	75
	.2041	87
	.1564	88
	.1005	94
	.0874	95
	.0540	99
	.0471	100
	.0261	104
	.0223	105
	.0112	109
	.0093	110
	.0052	113
	.0042	114
7 7 8	.5000	81
	.4764	82
	.2025	95
	.1860	96
	.1048	102
	.0941	103
	.0524	108
	.0461	109
	.0265	113
	.0229	114
	.0102	119
	.0086	120
	.0060	122
	.0049	123
7 8 8	.5111	88
	.4869	89
	.2076	103
	.1938	104
	.1039	111
	.0940	112
	.0544	117
	.0483	118
	.0254	123
	.0221	124
	.0104	129
	.0089	130
	.0053	133
	.0044	134

k = 3

n_1 n_2 n_3	α	$f(\alpha, 3, (n_1, n_2, n_3))$
8 8 8	.5104	96
	.4896	97
	.2087	112
	.1939	113
	.1085	120
	.0989	121
	.0537	127
	.0480	128
	.0263	133
	.0231	134
	.0115	139
	.0099	140
	.0053	144
	.0045	145

k = 4

n_1 n_2 n_3 n_4	α	$f(\alpha, 4, (n_1, \ldots, n_4))$
2 2 2 2	.5492	12
	.4508	13
	.2683	15
	.1929	16
	.1302	17
	.0829	18
	.0484	19
	.0262	20
	.0123	21
	.0052	22
	.0016	23
3 3 3 3	.5276	27
	.4724	28
	.2220	33
	.1823	34
	.1166	36
	.0907	37
	.0515	39
	.0374	40
	.0266	41
	.0183	42
	.0123	43
	.0080	44
	.0050	45

Table A.8 (continued)

k = 4

n_1 n_2 n_3 n_4	α	$f(\alpha, 4, (n_1, \ldots, n_4))$
4 4 4 4	.5183	48
	.4817	49
	.2172	57
	.1910	58
	.1058	62
	.0895	63
	.0514	66
	.0420	67
	.0272	69
	.0215	70
	.0130	72
	.0100	73
	.0056	75
	.0041	76
5 5 5 5	.5132	75
	.4868	76
	.2030	88
	.1846	89
	.1083	94
	.0962	95
	.0574	99
	.0498	100
	.0271	104
	.0230	105
	.0113	109
	.0093	110
	.0050	113
	.0040	114
6 6 6 6	.5101	108
	.4899	109
	.2004	125
	.1863	126
	.1052	133
	.0961	134
	.0529	140
	.0474	141
	.0265	146
	.0234	147
	.0104	153
	.0089	154
	.0057	157
	.0043	158

k = 5

n_1 n_2 n_3 n_4 n_5	α	$f(\alpha, 5, (n_1, \ldots, n_5))$
2 2 2 2 2	.5353	20
	.4647	21
	.2110	25
	.1625	26
	.1213	27
	.0878	28
	.0613	29
	.0412	30
	.0265	31
	.0162	32
	.0094	33
	.0051	34
	.0026	35
3 3 3 3 3	.5198	45
	.4802	46
	.2274	53
	.1982	54
	.1049	59
	.0874	60
	.0588	61
	.0475	62
	.0300	64
	.0234	65
	.0102	68
	.0076	69
	.0055	70
	.0039	71
4 4 4 4 4	.5131	80
	.4870	81
	.2059	93
	.1876	94
	.1113	99
	.0991	100
	.0521	105
	.0452	106
	.0288	109
	.0245	110
	.0102	115
	.0084	116
	.0056	118
	.0046	119

IAN C.T. NISBET, PH.D.
VICE PRESIDENT FOR LIFE SCIENCES
AND PATHOLOGY

CLEMENT ASSOCIATES, INC.
A SCIENCE AND ENGINEERING FIRM

1515 WILSON BOULEVARD
SUITE 700
ARLINGTON, VIRGINIA 22209
(703) 276-7700

MARY B. PAXTON, SC.M.
SCIENCE ADVISOR/BIOSTATISTICIAN

CLEMENT ASSOCIATES, INC.
A SCIENCE AND ENGINEERING FIRM
1515 WILSON BOULEVARD
ARLINGTON, VIRGINIA 22209

(703) 276-7700

Statistical Methods for Research Workers

by

SIR RONALD A. FISHER, Sc.D., F.R.S.

D.Sc. (Ames, Chicago, Harvard, London), LL.D. (Calcutta, Glasgow)

Fellow of Gonville and Caius College, Cambridge Foreign Associate United States National Academy of Science and Foreign Honorary Member American Academy of Arts and Sciences; Foreign Member of the Royal Swedish Academy of Sciences; Member of the Royal Danish Academy of Sciences; Foreign Member American Philosophical Society; formerly Galton Professor, University of London; formerly Arthur Balfour Professor of Genetics, University of Cambridge

THIRTEENTH EDITION—REVISED

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 39A WELBECK STREET, W.1
1958

11

First Published	1925
Second Edition	1928
Third Edition	1930
Fourth Edition	1932
Fifth Edition	1934
Sixth Edition	1936
Seventh Edition	1938
Eighth Edition	1941
Ninth Edition	1944
Tenth Edition	1946
Tenth Edition Reprinted	1948
Eleventh Edition	1950
Twelfth Edition	1954
Thirteenth Edition	1958

COPYRIGHT TRANSLATIONS

French
Presses Universitaires de France
108 Boulevard St Germain
Paris 6e, France

German
Oliver and Boyd Ltd
Tweeddale Court
Edinburgh, Scotland

Italian
Unione Tipografico
Editrice Torinese
Corso Raffaello 28
Torino, Italy

Spanish
Aguilar
S.A. de Ediciones
Juan Bravo 38
Madrid, Spain

Japanese
Messrs Sobunsha
129 1-Chome, Hinode-cho,
Toshima ku, Tokyo, Japan

Also Dover Publications. N.Y. expected 1958

BY THE SAME AUTHOR

THE GENETICAL THEORY OF NATURAL SELECTION, 1930, Oxford Univ. Press

THE DESIGN OF EXPERIMENTS, 1935, 1937, 1942, 1947, 1949, 1951. Oliver and Boyd Ltd.

STATISTICAL TABLES (With Frank Yates), 1938, 1943, 1948, 1953, 1957. Oliver and Boyd Ltd.

THE THEORY OF INBREEDING, 1949. Oliver and Boyd Ltd.

CONTRIBUTIONS TO MATHEMATICAL STATISTICS, 1949, John Wiley and Sons Inc., New York

STATISTICAL METHODS AND SCIENTIFIC INFERENCE, 1956, Oliver and Boyd Ltd.

PRINTED AND PUBLISHED IN GREAT BRITAIN BY
OLIVER AND BOYD LTD., EDINBURGH

EDITORS' PREFACE

THE increasing specialisation in biological inquiry has made it impossible for any one author to deal adequately with current advances in knowledge. It has become a matter of considerable difficulty for a research student to gain a correct idea of the present state of knowledge of a subject in which he himself is interested. To meet this situation the text-book is being supplemented by the monograph.

The aim of the present series is to provide authoritative accounts of what has been done in some of the diverse branches of biological investigation, and at the same time to give to those who have contributed notably to the development of a particular field of inquiry an opportunity of presenting the results of their researches, scattered throughout the scientific journals, in a more extended form, showing their relation to what has already been done and to problems that remain to be solved.

The present generation is witnessing "a return to the practice of older days when animal physiology was not yet divorced from morphology." Conspicuous progress is now being seen in the field of general physiology, of experimental biology, and in the application of biological principles to economic problems. Often the analysis of large masses of data by statistical methods is necessary, and the biological worker is continually encountering advanced statistical problems the adequate solutions of which approximate one, though validly applicable in an immense range of important cases. For other cases where the observations are measurements, instead of frequencies, it provides exact tests of significance. Of these the two most important are:—

(i) its use to test whether a sample from a normal distribution confirms or contradicts the variance which this distribution is expected on theoretical grounds to have, and

(ii) its use in combining the indications drawn from a number of independent tests of significance.

Ex. 14. *Agreement with expectation of normal variance.*—If $x_1, x_2, \ldots$, are a sample of a normal population, the standard deviation of which population is σ, then

$$\frac{1}{\sigma^2} S(x-\bar{x})^2$$

is distributed in random samples as is χ^2, taking n one less than the number in the sample. J. W. Bispham gives three series of experimental values of the partial correlation coefficient, each based on thirty observations of the values of three variates, which he assumes should be distributed so that $1/\sigma^2 = 29$, but which properly should have $1/\sigma^2 = 28$. The values of $S(x-\bar{x})^2$ for the three samples of 1000, 200, 100 respectively are, as judged from the grouped data,

$$35·0279, \quad 7·4573, \quad 3·6146,$$

whence the values of χ^2 on the two theories are those given in Table 21.

TABLE 21

	Exp. 1.	2.	3.	Total.	$\sqrt{2\chi^2}$.	Difference.
$29\,S(x-\bar{x})^2$ . .	1015·81	216·26	104·82	1336·89	51·71	+·79
$28\,S(x-\bar{x})^2$ . .	980·78	208·80	101·21	1290·79	50·81	−·11
Expectation (n) .	999	199	99	1297	50·92	

It will be seen that the true formula for the variance gives slightly the better agreement. That the difference is not significant may be seen from the last two columns. About 6000 observations would be needed to discriminate experimentally, with any certainty, between the two formulæ.

21·1. The Combination of Probabilities from Tests of Significance

When a number of quite independent tests of significance have been made, it sometimes happens that although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance. It is sometimes desired, taking account only of these probabilities, and not of the detailed composition of the data from which they are derived, which may be of very different kinds, to obtain a single test of the significance of the aggregate, based on the product of the probabilities individually observed.

The circumstance that the sum of a number of values of χ^2 is itself distributed in the χ^2 distribution with the appropriate number of degrees of freedom, may be made the basis of such a test. For in the particular case when $n = 2$, the natural logarithm of the probability is equal to $-\frac{1}{2}\chi^2$. If therefore we take the natural logarithm of a probability, change its sign and double it, we have the equivalent value of χ^2 for 2 degrees of freedom. Any number of such values may be added together, to give a composite test, using the Table of χ^2 to examine the significance of the result.

Ex. 14·1. *Significance of the product of a number of independent probabilities.*—Three tests of significance have yielded the probabilities ·145, ·263, ·087; test whether the aggregate of these three tests should be regarded as significant. We have

P	$-\log_e$ P	Degrees of Freedom.
·145	1·9310	2
·263	1·3356	2
·087	2·4419	2
	5·7085	6

$$\chi^2 = 11·4170$$

For 6 degrees of freedom we have found a value 11·417 for χ^2. The 5 per cent. value is 12·592 while the 10 per cent. value is 10·645. The probability of the aggregate of the three tests occurring by chance therefore exceeds ·05, and is not far from ·075.

In applying this method it will be noticed that we require to know from the individual tests not only whether they are or are not individually significant, but also, to two or three figure accuracy, what are the actual probabilities indicated. For this purpose it is convenient and sufficiently accurate for most purposes to interpolate in the table given (Table III), using the logarithms of the values of P shown. Either natural or common logarithms may equally be employed. We may exemplify the process by applying it to find the probability of χ^2 exceeding 11·417, when $n = 6$.

Our value of χ^2 exceeds the 10 per cent. point by ·772, while the 5 per cent. point exceeds the 10 per cent. point by 1·947; the fraction

$$\frac{·772}{1·947} = ·397.$$

The difference between the common logarithm of 5 and of 10 is ·3010, which multiplied by ·397 gives ·119; the negative logarithm of the required probability is thus found to be 1·119, and the probability to be ·076. For comparison, the value calculated by exact methods is ·07631.

22. Partition of χ^2 into its Components

Just as values of χ^2 may be aggregated together to make a more comprehensive test, so in some cases it is possible to separate the contributions to χ^2 made by the individual degrees of freedom, and so to test the separate components of a discrepancy.

Ex. 15. *Partition of observed discrepancies from Mendelian expectation.*—The table on p. 102 (de Winton and Bateson's data) gives the distribution of sixteen families of Primula in the eight classes obtained from a back-cross with the triple recessive.

The theoretical expectation is that the eight classes should appear in equal numbers, corresponding to the hypothesis that in each factor the allelomorphs occur with equal frequency, and that the three factors are unlinked. This expectation is fairly realised in the totals of the sixteen families, but the individual families are somewhat irregular. The values of χ^2 obtained by comparing each family with expectation are given in the lowest line. These values each correspond to 7 degrees of freedom, and it appears that in 6 cases out of 16, P is less than ·1, and of these structuring procedures. One of the major variables of the study was that of "psycho-therapeutic attraction." The basic data in Table 2 consist of the raw scores for this measure according to each of the four experimental conditions. Apply procedure (6), with the correction for ties given by (7).

2. Show directly, or illustrate by means of an example, that the maximum value of H is $H_{\max} = (N^3 - \sum_{j=1}^{k} n_j^3)/(N(N+1))$. For what rank configurations is this maximum achieved?

2. A DISTRIBUTION-FREE TEST FOR ORDERED ALTERNATIVES (JONCKHEERE, TERPSTRA)

Procedure. To test H_0 (2) against alternatives (see Comment 9) of the form

$$H_a: \tau_1 \le \tau_2 \le \cdots \le \tau_k, \tag{8}$$

where at least one of the inequalities is strict,

1. Compute $k(k-1)/2$ Mann-Whitney counts U_{uv}, $u < v$, where

$$U_{uv} = \sum_{i=1}^{n_u} \sum_{i'=1}^{n_v} \phi(X_{iu}, X_{i'v}) \tag{9}$$

and $\phi(a, b) = 1$ if $a < b$, 0 otherwise. That is, U_{uv} is the number of sample u before sample v precedences.

2. Let

$$J = \sum_{u<v}^{k} U_{uv} = \sum_{u=1}^{k-1} \sum_{v=u+1}^{k} U_{uv} \tag{10}$$

be the sum of these $k(k-1)/2$ Mann-Whitney counts.

3. At the α level of significance,

$$\begin{aligned} &\text{reject } H_0 \quad \text{if } J \ge j(\alpha, k, (n_1, \ldots, n_k)),\\ &\text{accept } H_0 \quad \text{if } J < j(\alpha, k, (n_1, \ldots, n_k)), \end{aligned} \tag{11}$$

where the constant $j(\alpha, k, (n_1, \ldots, n_k))$, which satisfies the equation $P_0\{J \ge j(\alpha, k, (n_1, \ldots, n_k))\} = \alpha$, is obtained from Table A.8.

Large Sample Approximation. Set

$$J^* = \frac{J - E_0(J)}{[\operatorname{var}_0(J)]^{1/2}} = \frac{J - \left\{\left(N^2 - \sum_{j=1}^{k} n_j^2\right)\Big/4\right\}}{\left\{\left[N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)\right]\Big/72\right\}^{1/2}}. \tag{12}$$

When H_0 is true, the statistic J^* has an asymptotic [min $(n_1, \ldots, n_k)$ tending to infinity] $N(0, 1)$ distribution. The approximate α-level test is

$$\begin{aligned} &\text{reject } H_0 \quad \text{if } J^* \ge z_{(\alpha)},\\ &\text{accept } H_0 \quad \text{if } J^* < z_{(\alpha)}. \end{aligned} \tag{13}$$

Ties. Replace $\phi(a, b)$ by $\phi^*(a, b) = 1$ if $a < b$, $\frac{1}{2}$ if $a = b$, 0 otherwise, so that for each between sample comparison where there is a tie, the contribution to the Mann-Whitney count will be $\frac{1}{2}$.

Example 2. Hundal (1969) described a study designed to assess the purely motivational effects of knowledge of performance in a repetitive industrial task. The task was to grind a metallic piece to a specified size and shape. Eighteen male workers were divided randomly into three groups. The subjects in the control group A received no information about their output, subjects in group B were given a rough estimate of their output, and subjects in group C were given accurate information about their output and could check it further by referring to a figure that was placed before them. The basic data in Table 3 consist of the numbers of pieces processed by each subject in the experimental period.

We apply Jonckheere's test with the notion that a deviation from H_0 is likely to be in the direction of increased output with increased degree of knowledge of performance.

From (9) we obtain

$$U_{12} = 22, \qquad U_{13} = 30.5, \qquad U_{23} = 26.5,$$

and from (10) we have

$$J = 22 + 30.5 + 26.5 = 79.$$

From Table A.8 we find $j(.0231, 3, (6, 6, 6)) = 79$ and thus using procedure (11) the lowest level at which we can reject H_0 (2) is .0231. Now let us apply the large sample approximation and compare it with the exact test. From (12) we compute

$$J^* = \frac{\{79 - (\frac{1}{4})[(18)^2 - (6^2 + 6^2 + 6^2)]\}}{\{[(18)^2(39) - 3(6)^2(15)]/72\}^{1/2}} = \frac{25}{12.37} = 2.02.$$

Table 3. Number of pieces processed

Control (no information)	Group B (rough information)	Group C (accurate information)
40 (5.5)*	38 (2.5)	48 (18)
35 (1)	40 (5.5)	40 (5.5)
38 (2.5)	47 (17)	45 (15)
43 (10.5)	44 (13)	43 (10.5)
44 (13)	40 (5.5)	46 (16)
41 (8)	42 (9)	44 (13)

Source. P. S. Hundal (1969).

* Although we do not need to perform the joint ranking to compute Jonckheere's statistic, we give these ranks here for later use in Sec. [illegible].3B.

Hence, using the approximate procedure (13), the lowest level at which we reject H_0 is .0217. Thus, for the task considered by Hundal, both the exact test and the large sample approximation indicate strong evidence of increased output with increase in degree of knowledge of performance.

Comments

9. In addition to degrees of knowledge of performance, other examples of ordered treatments are quality of materials, amount of practice, intensity of a stimulus, and temperature. Jonckheere's test should be preferred to the Kruskal-Wallis test (Section 1) when the treatments are ordered and the experimenter expects a deviation from H_0 to be in a particular direction. (If the direction expected is not the natural ordering $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$ of H_a, simply relabel the treatments so that the postulated order agrees with the natural order used here.) Note that the Kruskal-Wallis statistic does not utilize the partial prior information in a postulated alternative ordering. The statistic H (4) takes on the same value for all possible (k!) labelings of the treatments.

10. Consider J (10) and note that the term $\sum_{u<v}^{k} U_{uv}$ takes the postulated ordering into account. Consider, for simplicity, the case $k = 3$. Then $\sum_{u<v}^{3} U_{uv} = U_{12} + U_{13} + U_{23}$, and if $\tau_1 < \tau_2 < \tau_3$, U_{12} would tend to be larger than $n_1n_2/2$ (its null expectation); U_{13} would tend to be larger than $n_1n_3/2$; U_{23} would tend to be larger than $n_2n_3/2$ and, consequently, $J = U_{12} + U_{13} + U_{23}$ would tend to be larger than its null expectation $(n_1n_2 + n_1n_3 + n_2n_3)/2 = \{[N^2 - (n_1^2 + n_2^2 + n_3^2)]/4\}$. This serves as partial motivation for the J test.

11. A little thought will convince the reader that J can be computed from the joint ranking of all $N = \sum_{j=1}^{k} n_j$ observations. That is, although we do not need to perform this joint ranking in order to compute J, given the ranking we can, without knowledge of the actual X_{ij} values, retrieve the value of J. Thus one way to obtain the null distribution of J is to follow the method of Comment 5; namely, use the fact that under H_0 (2) all $N!/(\prod_{j=1}^{k} n_j!)$ rank assignments are equally likely, and compute the associated value of J for each possible ranking. Consider how this would work in the small sample size case of $k = 3$, $n_1 = n_2 = n_3 = 2$, which was used in Comment 5. We can easily calculate the value of J for each of the 15 rank configurations displayed in Comment 5. However, we cannot eliminate the calculations corresponding to the other 75 rank configurations (as we were able to do in the case of the Kruskal-Wallis H statistic), since J does depend on the particular numbering of the treatments. For example, consider rank configurations (a) and (a^*):

(a)

I	II	III
1	3	5
2	4	6

(a^*)

I	II	III
3	1	5
4	2	6.

Note that (a) and (a^*) may be viewed as the same except that we have interchanged the roles of treatments I and II. The J value for (a) is readily found to be $J = 12$, whereas for (a^*) we have $J = 8$. [The statistic H (4) takes on the same value, 4.57, for (a) and (a^*).] Without resorting to complete enumeration, we can check a value in Table A.8 as follows. Configuration (a) will yield the largest possible value of J, namely, $J = 12$, and no other configuration will yield a value that large. Thus $P_0(J \geq 12) = P_0(J = 12) = \frac{1}{90} = .0111$, which agrees with the appropriate entry in Table A.8.

12. Table A.8 gives critical values for $n_1 \leq n_2 \leq n_3$ situations. However, critical points for (n_1, n_2, n_3) configurations not in this order can be obtained by simply putting the three sample sizes in increasing order and then entering Table A.8. (This is a consequence of certain symmetry properties of the distribution of J.) Thus, for example, to find critical values for the case $n_1 = 4$, $n_2 = 6$, $n_3 = 2$, enter Table A.8 at $n_1 = 2$, $n_2 = 4$, $n_3 = 6$.

13. For $k = 2$, the procedure defined by (11) reduces to the one-sided Mann-Whitney-Wilcoxon test (Section 4.1).

Properties

1. Consistency: The condition n_j/N tends to λ_j, $0 < \lambda_j < 1$, $j = 1, \ldots, k$, is sufficient to insure that the test defined by (11) is consistent against the H_a (8) alternatives. For a more general consistency statement, see Terpstra (1952).

2. Efficiency: See Puri (1965) and Section 5.

REFERENCES. The test based on the J statistic was proposed by Terpstra (1952) and independently by Jonckheere (1954a). The first generalization of the Mann-Whitney-Wilcoxon two-sample test, with ordered alternatives in mind, was given by Whitney (1951). Whitney treated the case $k = 3$, and his procedures are not equivalent to the J test when the latter is specialized to $k = 3$. Chacko (1963) proposed a rank analog of a normal theory ordered alternatives test developed by Bartholomew (1959a, 1959b, 1961a, 1961b). Puri (1965) generalized Jonckheere's test to a class of tests including a normal scores analog of the J test. Further generalizations were given by Tryon and Hettmansperger (1971). For a different approach to the utilization of partial prior information, see Abelson and Tukey (1963).

PROBLEMS

3. Apply Jonckheere's test to the data of Table 2 using the postulated ordering $\tau_1 < \tau_2 < \tau_3 < \tau_4$.

4. The statistic J can be computed either from (a) the joint ranking of $N = \sum_{j=1}^{k} n_j$ observations or from (b) $k(k-1)/2$ "two-sample" rankings. Explain.